Analysis Of Personality Assessment Based On The Five-Factor Model Through Machine Learning

Noureen Aslam, Khalid Masood Khan, Afrozah Nadeem,Sundus Munir, and JavairyaNadeem

Abstract: Social media is one of the most popular platformsand people from all the diverse fields such as students or professionals explore social media daily. It is a platform where people are available from different cultures and religions. With the advancement of technologies in every field of life, there is an increased demand for social media. Whenever people go online they generate rich data through their smartphones or internet pads. Their texting style, taste in music, books, likes, dislikes, sharing posts reveal their personality, therefore social media is an ideal platform to study the human personality. Personality has been considered as an essential factor and it is a combination of different attributes that make a person unique from one another. In our proposed work, we used Twitter data and myPersonality datasets to perform an objective assessment using a deep sequential neural network and multi-target regression model for predicting personality traits. The proposed algorithm is based on the Five-Factor Model (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism). The efficacy of the proposed technique has been measured by MSE, MAE, Precision, Recall, and F1-Score. Experimental results show that our model is robust and ithas outperformed the existing techniques to predict human personality traits.

Index Terms: Deep sequential neural network, Five-Factor model, Multi-target regression model, Machine learning, Social media, Twitter.

1. INTRODUCTION

1.1 The Five-Factor Model

The Five-Factor model is one of the well-established models to recognize personality. It uses words to identify personality and analyze in which trait a person fits. It characterizes a person into five traits i-e agreeableness, conscientiousness, extraversion, neuroticism, and openness [1].

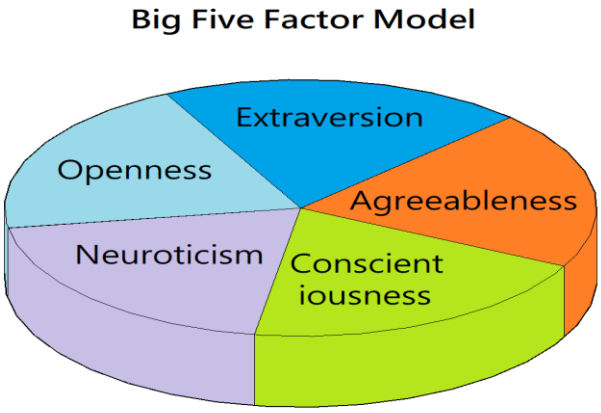

***Fig. 1**. Five-Factor Model attributes*

The Five-Factor model is also called the “BIG5” or “Ocean” model. According to Hirschfeld, the Five-Factor Model provides the prestigious dimensions of personality and the five traits of “FFM” or “Big Five” can be stated below:

O- Openness:Openness is a dimension of the Five-Factor Model. Openness relates to an individual who is creative, imaginative, artistic, understanding, curious, politically liberal, traditional, competitors, and love to travel new places. They are successful if they pursue the field of an accountant, auditor, judge, and financial manager.

C- Conscientiousness:Conscientiousness is another dimension or trait of the Five-Factor Model which has two basic features dependability and accomplishment. The highly conscientious people are well planned, organized, dutiful, reliable, purposeful, impulse control, workaholic, self-disciplined, determined, and confident. They tend to be less bound by plans and rules but more tolerant.

E- Extraversion:Extroversion is a classic dimension of the Five-Factor Model. The highly extroverted people are affectionate, friendly, excitement seeker, energetic, assertive, optimistic, outgoing, charismatic, and talkative. Most of the extrovert gain energy from their surroundings. They become successful in the future if they pursue politics and sales as their career.

A- Agreeableness:Agreeableness is a dimension that is concerned with the nature of one's associations with others. The highly agreeable people are cooperative, courteous, friendly, trustworthy, kind, tolerant, pardoning, polite nature, peacekeeper, calm, imaginative, caring, and raise new ideas. They ignore their needs for others, they are good team worker but don't work as a leader. They are easily influenced, adopt group opinions, working in the background, and keeping a positive relationship with others.

N- Neuroticism:Neuroticism is another classic dimension of the Five-Factor Model and they are reversely referred to as “Emotional Stability”. The individuals who have a higher degree of neuroticism are equivalent to emotional uncertainty, pessimistic, sensitive, insecure, unstable, nervous, easily depressed, vulnerable, irritated, shocked easily, moody, experience negative emotions like (anxiety, anger, and depression) and never satisfied with their lives. They pursue pilot, engineering, and manager as their career.

- *Noureen Aslam is currently pursuing an MPhil degree program in computer science from Lahore Garrison University, Pakistan, PH-03132138097. E-mail: noureenaslam2@gmail.com*
- *Dr. Khalid Masood Khan is an Associate professor at Lahore Garrison University, Pakistan, PH-03357400418, E-mail: Khalid.masood@lgu.edu.pk*
- *Afrozah Nadeem is a lecturer at Lahore Garrison University, Pakistan, PH-03004256228, Email: afrozah@lgu.edu.pk*
- *Sundus Munir is an Assistant professor at Lahore Garrison University, Pakistan, PH-03238437975, Email: sundusmunir@lgu.edu.pk*
- *JavairyaNadeemis astudent of MSCS at Lahore Garrison University, Pakistan, PH-03364206694. E-mail: javairyakhan@gmail.com*

Table 1.
Low vs High scores of the Five-Factor model attributes
"FIVE-FACTOR" or "OCEAN" Model Traits

O	OPENNESS	O refers to openness. The ones who curious to learn new and interesting things and have a broad range of interests. Never hesitate to experience new things.	High Degree	curious, imaginative, love traveling, creative or artistic, music lover, original, intellect, thoughtfulness, revolution lover, aware of their feelings, deficiency of focus and love new experiences.
			Low Degree	closed-minded, traditional, closed to experience, a good planner or execute their plans efficiently and strictly follow the rules and regulations.
C	CONSCIENTIOUSNESS	C refers to Conscientiousness . The ones who do their work timely and properly in an organized way.	High Degree	well planned, organized, dutiful, reliable, purposeful, impulse control, workaholic, self-disciplined, determined, and confident.
			Low Degree	unorganized, spontaneous, careless, flexible, timid, lazy, and relaxed.
E	EXTRAVERSION	E refers to Extraversion. The ones who get energy from others by interacting with them.	High Degree	energetic, affectionate, friendly, excitement seeker, assertive, optimistic, outgoing, charismatic, talkative, and sociable.
			Low Degree	reserved, low key, quiet, unsociable, less energetic, deliberate, shy and expand more time alone.
A	AGREEABLENESS	A refers to Agreeableness. The ones who trust others. Acceptand compromise their ideas with others. They are soft-hearted people.	High Degree	cooperative, courteous, friendly, trustworthy, kind, tolerant, forgiving, polite nature, peacekeeper, calm, imaginative, agreeable, loving, and appreciative.
			Low Degree	expressive, aggressive,harsh,analytical , born leaders, rude, self-centered, and challenge acceptors.
N	NEUROTICISM	N refers to Neuroticism. The ones who are emotionally unstable and experience negative feelings like nervousness, irritation, sadness, jealousy, fury, and depression.	High Degree	Insecure, nervous, moody, irritated, depressed, shocked, upset, and struggles to bounce back after traumatic events.
			Low Degree	emotionally stable, optimistic, calm, composed, peace lovers, easily recover from depression, satisfied with their lives, and discover the best out of the worst.

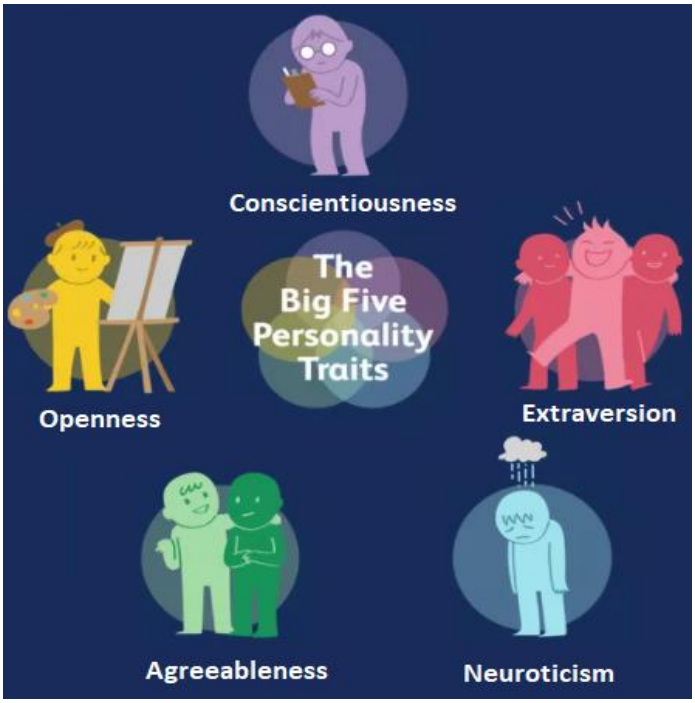

Fig. 2*.Emotions of the Big Five personality traits*

The four main researchers J. M. Digman, A. Comrey, Lewis Goldberg, and Dr. Naomi Takemoto-Chowk studied the existing personality test in a seminar that was conducted in Honolulu in 1981. They clinched that Norman's major five factors are the most promising traits of human personality. In the 1980s, the five-factor model was widely accepted by personality researchers. Peter Saville and his co-workers combined the five-factor model with OPQ i-e occupational, personality, and Questionnaire, and in 1985 McCrae and Costa followed the five-factor model as NEO i-e Neuroticism, Extraversion, and Openness [2]. Every individual is different from one another due to their unique psychological structure (thinking, behavior, and preferences). It is impossible to completely compare one person to another because every individual has specific traits of personality. It is true some traits resemble one another but not a complete personality resembles others. People are different from each other along with a similar set of traits [3]. It is a general desire to understand the personality of a person and a lot of research has beenperformed on the personality assessment. There are several models proposed to assess personalities like "MBTI (Mayers-Briggs Type Indicator)" proposed by Katharine Cook Briggs and Isabel Briggs Mayer initiated by the Carl Jung theory, "Freud's Theory" developed by Sigmund Freud and he believed that there are three facets of the mind (id, ego, and superego) that contain an individual's personality as well as he stated that there is a relation between nurture and nature, "Eysenck's Personality Theory" developed by Eysenck in 1947 and he believed that extraversion, neuroticism, and psychoticism are basic traits of personality that reveal about personality, "Cattell's Theory" developed by Cattell in 1965, he revealed that there are 16 personality traits and completely disagreed with Eysenck theory but The Five-Factor model considered the best model for personality assessment because the five characteristics of the Five-Factor Model i-e Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism are fundamental

characteristics that build an individual's overall personality [4]. The Five-Factor model is one of the reliable, predictive, and efficient personality assessment models [5]. Nowadays with the advancement of the technical world, people mostly communicate on social media sites. The Digital generation spending more time on social media and every individual has their account on social networking sites like Twitter, Facebook, Instagram, Whatsapp, and the list goes on and on through which they communicate with each other. Whenever people go online they generate data through social media or by using their cellphones. The language that is used by People on social media is full of psychological content by using it generates a valid and fast personality assessment [6].In our research, we used users generated content to insight the personality traits of users without having them fill out any questionnaire. We are interested in words used by the users on their profiles to predict their personality traits. We used two different datasets for our work: a myPersonality dataset collected from Facebook (labeled data) and a Twitter dataset collected from user's Twitter profiles (extract 3200 tweets of a single user to predict their personality traits). The research study is presented as follows. Section 2 demonstrates the related work. Section 3 explains the proposed methodology. Section 4 defines the evaluation metrics. Section 5 demonstrates the results of our proposed model. A comparison of different classification models is discussed in section 6. Conclusion and Future work are presented in the last section.

2 Related Work

Every person has a different pattern of feeling, thinking, and behavior that makes one's personality unique from one another. Numerous amount of research has been done in the field of personality assessment by adopting different classification methods, Datasets, and personality models. Given below is a brief overview of related work. In 2020, Xiao Kun Wu et al. assessed personality through Twitter tweets of social media users using a weighted Random Forest (weighted-RF) classifier. They used two datasets about civil rights events and compared weighted-RF with seven different machine learning algorithms as well as claimed that weighted Random Forest performed the best among all the other classifiers [7]. Rohit GV et al. predict personality using Facebook status that can be shared by users on their profiles. They used the BIG5 model and Random Forest Classifier for their research and achieved 64.25% accuracy or 5.25 mean square error was achieved by using a random forest regressor [8]. In 2019, Aditi V. Kunte et al. used a real-time Twitter dataset for the personality prediction task. They used three different machine learning algorithms (Linear Discriminate Analysis, Adaboost, and Multinomial Naive Bayes). After comparing these three algorithms they concluded thatMultinomial Naive Bayes got the highest accuracy that is 73.43% [9].In 2018, Giulio Carducci et al. Presented a supervised learning approach to identify personality traits through an individual's tweets. They used myPersonality and Twitter datasets for their work and word embedding used as a vectorizer. SVM (Support Vector Machine), Linear regression, and Lasso classification models are used and concluded that linear regression performs the worse [10]. K. Kircaburun and M. D. Griffiths explored the associations among personality, daily usage of the internet, self-seeking, and Instagram addiction. The sample size is N=752 these are university students who completed a survey report that contains FFM inventory, Instagram addiction, and self-seeking scale. They concluded that daily usage of the internet was negatively related to Instagram addiction either self-seeking, consciousness, and agreeableness are three factors that were negatively related to Instagram addiction [11]. In 2017, Varshney et al. recognized human personality via analyzing their handwriting. For the analysis of the handwriting-based system, they used the following terms: margin, spacing, connections, slant, letter size, speed, clarity, pressure, zones, a large middle zone, small middle zone, and upper zone extension [12]. Gupta et al. contribute to the personality prediction task. For their research, they collected Facebook profile data like users name, date of birth, gender, age, qualification, marital status as well as Facebook activities like dislike, and posts to predict personality traits. The activities performed by every individual is essential because it reveals valuable insight to predict the user's concern, behavior, nature, and sentiments. They used the k nearest neighbor algorithm to train their dataset [13]. Ahmad et al. predict the personality traits of a Twitter user by using the DISC model. They extracted about one million recent tweets from Twitter and divide their methodology into three main steps that are data extraction, analysis, and visualization. They extracted the data using Rapidminer which is one of the best machine learning tools and offers multiple text processing packages. After extraction, analysis, and visualization they predict the personality traits of users [5].In 2015 Yakasai et al. examine the effect of FFM attributes on the performance of salespersons. There are three main objectives of their research first of all they analyze the connection among the attributes of FFM, salesperson performance, and salesperson purchaser orientation. Secondly, they explored the impact of culture on the proposed model, and Finally, they proposed a model on the performance of a salesperson as well as include the impact of client orientation and culture on sales performance [14]. In 2014, Dehghanan et al. examine the impacts of the Five-Factor model attributes on emotional intelligence. For this, they worked on a few Iranian organizations and used two questionnaires “McCrae and Costa questionnaire (NEO-PI-R)” to measure the personality attributes and “Bradbury and Grieve questionnaire” to measure the impacts of emotional intelligence. They considered the five hypotheses by using regression and structural equation modeling and concluded that Openness, Consciousness, Extraversion, and Agreeableness affects positively on emotional intelligence whereas Neuroticism affectsnegatively [15].In 2013, Alam et al. contribute to the field of personality prediction and compare the performance of different classification methods also concluded that MNB (Multinomial Naive Bayes) performs better as compare to SMO (Sequential minimal optimization) and BLR (Bayesian Logistic Regression) [16]. In 2011, Golbeck et al. collected 2000 recent tweets of users and perform text analysis. They choose the BIG 5 model and analyze tweets text through two methods LIWC (Linguistic Inquiry and Word Count) and MRC psycholinguistic database. When the features are obtained, they apply two different machine learning algorithms for personality prediction (Zero and Gaussian process) [17].

The scope of this paper is to predict the traits of human personality. Personality prediction can be used in a candidate's hiring process and also helps to identify criminals or negative-minded people.

3 Methodology

The methodology followed in this research is presented in figure 3, its objectives are to identify the personality scores of an individual by evaluating their online social media data and create a model for personality prediction. To build this model several modules are used such as Netminer, deep sequential neural network, and multi-target regression model.

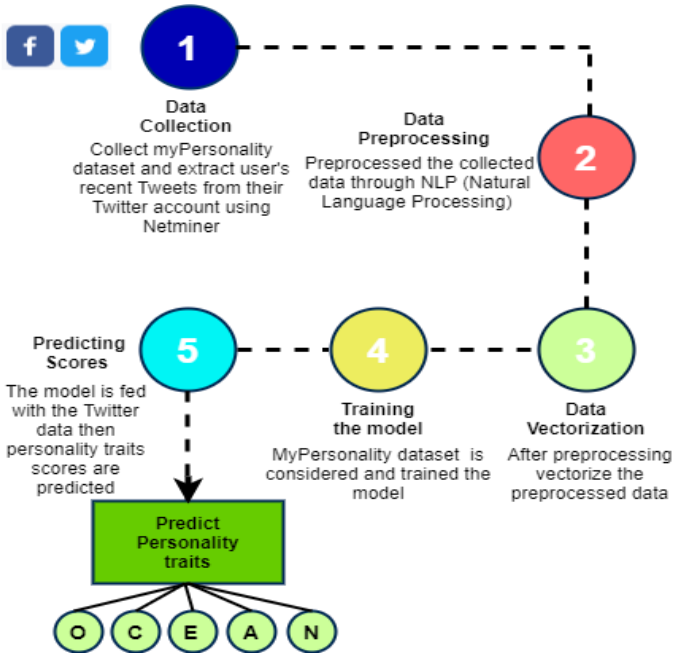

***Fig. 3**.Proposed System Architecture*

3.1 Data Collection or Datasets

For this study, two datasets are used 1) myPersonality dataset 2) Twitter dataset.

1. myPersonality Dataset

The myPersonality project was a Facebook application. In June 2007 David Stillwell has proposed it. At that time he was a student of Nottingham University and now a lecturer of Cambridge University and Michal Kosinski joined David Stillwell in 2009, at that time he was a student of Cambridge University and now a lecturer of Stanford University. In January 2012 it was closed due to a deficiency of time to continue it. During that period approximately 7.5 million users completed measures on my personality project. According to the official website of my personality project, there are more than 50 chapters and articles that have been published on it. It is free of cost and provides quality data [18]. It allows Facebook users to access 25 phycological tests in which also include the Five-Factor model questionnaires.

Table 2.
Facebook Dataset details

Dimensions	Totals
Facebook Users	9,917
Statusbefore preprocessing	9,917
Status after preprocessing	9,821
Words before preprocessing	14,4,021
Words after preprocessing	7,9755

2. Twitter Dataset

For the Twitter dataset, we used the netminer tool to extract user's tweets. We extracted 3200 recent tweets of random Twitter users as well as extracted 3200 recent tweets of Imran khan and Bill gates to predict their personality traits. The figure below is a snapshot of Twitter data which contains nine columns i-e (tweet_id, created_time,twitter_screen, app, lang, hashtag, retweet_id, full_text, and tweetClass).

	tweet_id	created_time	r_screen_na	app	lang	hashta
1	13000556	2020-08-30 22:(	min_ciencia	Twitter	es	None
2	13001009	2020-08-31 01:(	mphermosill	Twitter	es	COVID
3	13001009	2020-08-31 01:(	71nuevaespa	Twitter	es	EnVide
4	13001009	2020-08-31 01:(	EmergenciaA	Twitter	es	corona
5	13001009	2020-08-31 01:(	dailywildcat	Buffer	en	COVID
6	13000999	2020-08-31 00:(	LiveLawIndia	Twitter	en	NEETJ
7	13001009	2020-08-31 01:(	RishabTej2	Twitter	en	NEETJ
8	13001009	2020-08-31 01:(	AmerMedica	Sprinkl	en	None
9	13001009	2020-08-31 01:(	ALaMagdale	Twitter	es	LaMag

***Fig. 4**. Twitter Dataset snapshot*

Table 3.
Twitter Dataset details

Dimensions	Totals
Twitter Users	3200
Tweets before preprocessing	3,20,000
Tweets after preprocessing	3,19,550
Words before preprocessing	48,54,413
Words after preprocessing	33,94,488

3.2 Data pre-processing

All the textual data has been collected and save in the ".csv" file then undergo the next phase which is preprocessing for further proceedings. Preprocess the data through NLP(Natural Language processing). In the data preprocessing phase we did the following steps:

1. Eliminates all data points with Null/Nan values.
2. Remove all punctuation marks, special characters, and symbols.
3. Case conversion, to lower case.
4. Stopwords removal. Stop words are the most commonly used words (i.e. "is", "the", "what" etc.). They are useless, so we eliminate these all.
5. Lemmatization, the purpose of lemmatization is grouping the words. Suppose there are words like enjoy, enjoyed, enjoying through lemmatization it should be considered as a single item.

Original Text	Normalized text
likes the sound of thunder.	like sound thunder
is so sleepy it's not even funny that's she can't get to sleep.	sleepy even funny get sleep
is sore and wants the knot of muscles at the base of her neck to stop hurting. On the other hand, YAY I'M IN ILLINOIS! <3	sore want knot muscle base neck stop hurting hand yay illinois
likes how the day sounds in this new song.	like day sound new song
is home. <3	home
www.thejokerblogs.com	www thejokerblogs com
saw a nun zombie, and liked it. Also, *PROPNAME* + Tentacle!Man + Psychic Powers = GREAT Party.	saw nun zombie liked also propname tentacle man psychic power great party
is in Kentucky. 421 miles into her 1100-mile journey home.	kentucky mile mile journey home
was about to finish a digital painting before her tablet went haywire. Is now contemplating the many ways she wishes to exact her revenge on faulty technology.	finish digital painting tablet went haywire contemplating many way wish exact revenge faulty technology
is celebrating her new haircut by listening to swinger music and generally looking like a doofus.	celebrating new haircut listening swinger music generally looking like doofus
has a crush on the Green Lantern.	crush green lantern

Fig. 5*.Data before and after pre-processing*

3.3 Data vectorization

Once the data has been cleaned by eliminating null values, punctuation marks, special characters, symbols, and stop words then proceed to the next phase which is data vectorization. TF-IDF (Term frequency-inverse Document Frequency) vectorizer is used to convert the textual data into vectors. It assigns the frequency to words like how many times a single word can appear in a file/document as well as downscales word that appears a lot of time in a document. The TF (Term Frequency) can be computed by using the formula that is mentioned below:

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k ni,j} \quad (1)$$

Each file has its own TF (Term Frequency), which is computed by dividing a word that appears the number of times in a file by the total number of words. The IDF (Inverse Data Frequency) can be computed by using the formula that is mentioned below:

$$\mathrm{idf}_{i,j=} = \log(\frac{N}{df_i}) \quad (2)$$

Through IDF weight/frequency of a word can be determined by a log of the number of files divided by the number of files that have the word “a”. Finally, TF-IDF can be calculated by simply multiplying TF by IDF that is mentioned below:

$$\mathrm{tfidf}_{i,j=}\ \mathrm{tf}_{i,j} \times \log(\frac{N}{df_i}) \quad (3)$$

So, by this method, we identify the frequency/weight of all the words in a document.

3.4 Model Training and Testing

In this phase, we consider myPersonality dataset and split the data into training and testing(80% for training and 20% for testing). During this research, several machine learning algorithms are studied to identify the ideal model. Therefore, deep sequential neural networks and multi-target regression models are chosen for our work. Using both deep sequential neural networks and multi-target regression model prediction of the OCEAN or FFM traits is successful. Training and testing of the proposed model have been done on the myPersonality dataset. The model is trained using deep sequential neural networks and a multi-target regression model then a trained model is implemented on the Twitter dataset to predict Twitter user's personality traits.

4 Evaluation Metrics

The performance of the system has been measured by evaluation metrics in which include MSE (Mean square Error), MAE (Mean Absolute Error), Precision, Recall, F1-score, and average precision. The following are the equations that can be used to compute all these mentioned evaluation metrics:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \quad (4)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}_i| \quad (5)$$

Where,
$\hat{Y}_i$= Predicted Value,
Y_i= Original value,
N= Number of observations

$$\mathrm{Precision} = \frac{\mathrm{True\ Positive}}{\mathrm{True\ Positive} + \mathrm{False\ Positive}} \quad (6)$$

Precision is the fraction of related occurrences among the retrieve occurrences and it is also known as positive predictive value. It is used to find out when the cost of a False Positive is high and precision is the ability of the classifier that doesn’t label negative samples as positive.

$$\mathrm{Recall} = \frac{\mathrm{True\ Positive}}{\mathrm{True\ Positive} + \mathrm{False\ Negative}} \quad (7)$$

The recall is the fraction of correctly identified occurrences over the total amount of relevant occurrences and it is also known as sensitivity.

$$\mathrm{F1-Score} = 2\ \mathrm{x}\ \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (8)$$

5 Experiment and Results

In this section, We discuss the performance and results of the experiments which we carried out for predicting personality traits of an individual based on his social media data. Jupyter Notebook is used as an IDE (Integrated Development Environment) and python is used as a programming language. The Netminer tool is used to extract Twitter data. Multi-target regression and deep sequential neural networks are the models used. Table 4 shows the MSE (Mean Square Error) of the proposed model and the overall combined MSE is 0.20. Table 5 shows the MAE (Mean Absolute Error) of the proposed model and the overall combined MAE is 0.34. Table 6 shows Precision (0.72), Recall (0.79), F1-score (0.72), training accuracy (94%), and testing accuracy (78%).

Table 4.
Proposed model MSE

Personality Traits	MSE of Training Data (Y)	MSE of Testing Data ($\hat{Y}$)	Mean Square Error $(Y - \hat{Y})^2$

MSE- Ext	0.24436	0.89839	0.42775
MSE- Neu	0.19388	0.68939	0.24553
MSE- Agr	0.182681	0.58182	0.15936
MSE- Con	0.201085	0.641281	0.19377
MSE- Opn	0.141151	0.428170	0.08236
Combined	0.192633	0.647813	0.20720

Table 5.
Proposed model MAE

Personality Traits	MAE of Training Data	MAE of Testing Data	Mean Absolute Error | Y – Ŷ|
MAE- Ext	0.32268	0.75396	0.4313
MAE- Neu	0.28015	0.66275	0.3826
MAE- Agr	0.28680	0.61667	0.3298
MAE- Con	0.29972	0.63555	0.3358
MAE- Opn	0.24962	0.48706	0.2374
Combined	0.28779	0.63120	0.3434

Table 6.
Proposed model evaluation metrics

Evaluation Metrics	O	C	E	A	N	AVG
Precision	0.74	0.82	0.7	0.78	0.56	0.72
Recall	0.99	0.77	0.8	0.64	0.78	0.79
F1-Score	0.78	0.75	0.67	0.68	0.72	0.72
Training Accuracy	0.95	0.94	0.94	0.94	0.94	0.94
Testing Accuracy	0.71	0.80	0.78	0.75	0.88	0.78

The above tables report the accuracy of the proposed model which is best among all the other machine learning classifiers that can be demonstrated in section 6. Since we have multiple outputs so we build a multi-label personality model. Our proposed model is consists of a fully connected neural network with 5 output neurons, each neuron predicts scores for one of each personality trait according to the BIG 5 model. The model categorizes the words according to the BIG 5 model. For example, frustrated, anxiety, anger are the words that relate to neurotics so the proposedalgorithm categorized the person into five factors that are openness, Consciousness, extraversion,agreeableness, and neuroticism. The model predicts thescores that how much percent a person extrovert, conscious, open, agreeable, and neurotic. The figure given below shows the scores for each personality trait of Facebook data based on the BIG 5 model.

STATUS	sEXT	sNEU	sAGR	sCON	cOPN
Facebook me marea. Me hates it long time T-T	2.45	4	2.85	2.35	4.1
It's a beautiful day in a neighborhood, a beautiful day in a neighborhood, somewhere else in the world...	2.75	3.25	2.5	4	4.65
About mornings and winter,and magic.	2.15	2.15	4.1	2.9	4.6
little things give you away.	2.15	2.15	4.1	2.9	4.6
is wishing it was Saturday.	4.05	3.35	3.8	3.95	4.5
is studying hard for the G.R.E.	4.05	3.35	3.8	3.95	4.5
snipers get more head	1.4	4.05	3.3	3.4	3.95

Fig. 6*.Facebook data personality scores*

Id	EXT	NEU	AGR	CON	OPN
100c885443c4d	330.11023	279.57645	385.7268	311.04608	394.6828
1017647e21e7e	328.54993	267.64178	350.7902	329.62323	398.63647
1023e1e534622	327.1196	295.08768	361.0858	328.0525	396.74408
10283d7f37d33	340.9468	271.3129	380.28284	346.2448	401.89316
1069f66c9d5862	322.32977	283.9491	374.62717	330.1901	404.35574
1075db9bdc0e9	322.8298	290.5713	378.05905	323.59387	415.9522
1088a2d04fa518	341.31573	254.23639	393.92157	353.38324	389.86414
109339c7630fb4	331.48013	263.47214	369.8795	343.95514	421.62564
10fa401266a889	321.0616	273.63647	368.746	347.57883	409.4187
110717a25ceca	322.2838	297.37878	378.4646	298.3721	381.3598
1120cf1015b427	335.27274	280.10043	385.3573	310.97458	397.91937
1127a028ad975	325.53757	286.15454	389.62692	315.7407	390.50577
112e50b50d4d7	330.3335	285.35156	365.9688	325.1127	403.84808
1148bb6a6c161	326.33234	271.04837	380.96838	320.12347	389.09665
115f9a92c9f7f15	332.30954	271.8574	360.70615	348.38925	410.64355
116651423c0b0	339.8267	270.48203	369.8952	336.2975	411.04837

Fig. 7*.Twitter data personality scores*

According to the proposed model prediction, Twitter users have the highest scores in openness and the lowest scores in Neuroticism. We also predict the personality traits of Mr. Imran Khan and Bill gates that's why we extracted the recent tweets of Mr. Imran khan and Bill gates to estimate their personality traits.

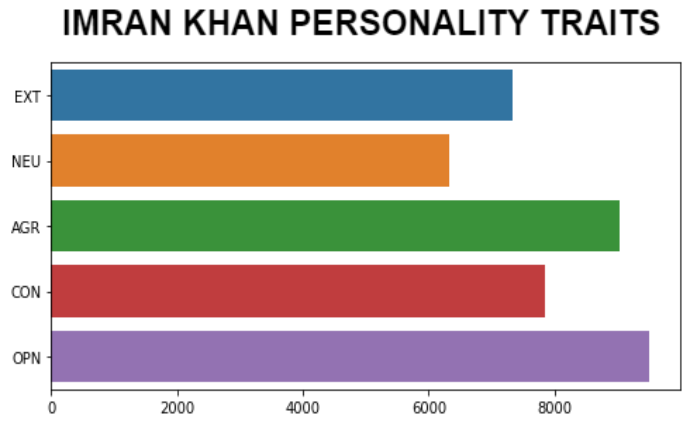

***Fig. 8**. Mr. Imran khan's personality traits*

The above figure shows Imran Khan's personality traits. Imran Khan has the highest scores in openness and agreeableness and the lowest scores in Neuroticism and Extraversion.

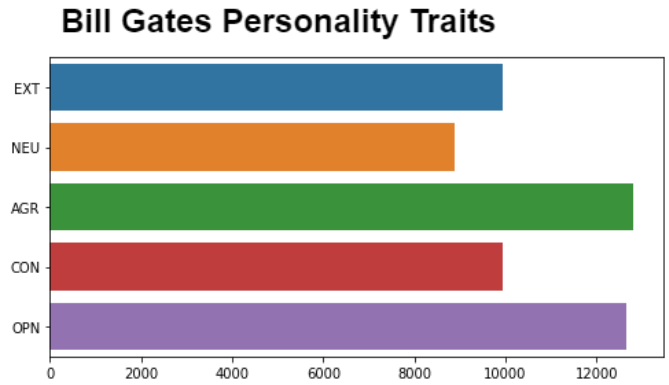

***Fig. 9**. Mr. Bill Gates's personality traits*

6 Comparison

To prove the accuracy, we compare the mean squarer error of different machine learning algorithms i-e Multinomial naive Bayes, logistic regression, Gaussian Naive Bayes with the proposed model. Our proposed model has the least mean square error. So, its performance is much better than other machine learning algorithms. Table 7 compares the results of the mean square error of Multinomial Naive Bayes, Logistic regression, and Gaussian Naive Bayes with the proposed model.

Table 7.

MSE comparison of different machine learning algorithms

Personality Traits	Multinomial Naive Bayes (MSE)	Logistic regression (MSE)	Gaussian Naive Bayes (MSE)	"Proposed model" SNN Multi-target regression (MSE)
cOPN	0.3915	0.3683	0.3785	0.42775
cCON	0.412	0.4053	0.4214	0.24553
cEXT	0.3986	0.3833	0.3931	0.15936
cAGR	0.4124	0.3841	0.4037	0.19377
cNEU	0.2719	0.2483	0.2455	0.08236

Table 8.

Evaluation metrics comparison of different machine learning algorithms

Classification Models	Evaluation Metrics	Personality Traits					
		EXT	NEU	AGR	CON	OPN	AVG
1. Logistic Regression	Precision	0.6	0.67	0.58	0.62	0.74	0.64
	Recall	0.6	0.68	0.59	0.62	0.77	0.65
	F1-Score	0.57	0.62	0.58	0.61	0.68	0.61
	Training Accuracy	0.78	0.76	0.80	0.79	0.76	0.77
	Testing Accuracy	0.59	0.67	0.58	0.61	0.77	0.64
2. Multinomial Naive Bayes	Precision	0.59	0.67	0.59	0.62	0.72	0.63
	Recall	0.59	0.67	0.59	0.62	0.77	0.64
	F1-Score	0.55	0.60	0.58	0.61	0.68	0.60
	Training Accuracy	0.79	0.74	0.80	0.80	0.77	0.78
	Testing Accuracy	0.590	0.67	0.59	0.61	0.76	0.64
3. Gaussian Naive Bayes	Precision	0.60	0.60	0.57	0.58	0.68	0.60
	Recall	0.55	0.50	0.56	0.54	0.52	0.53
	F1-Score	0.53	0.50	0.56	0.53	0.55	0.53
	Training Accuracy	0.721	0.699	0.82	0.74	0.68	0.73
	Testing Accuracy	0.543	0.50	0.561	0.544	0.519	0.53
4. Random Forest	Precision	0.55	0.60	0.54	0.57	0.70	0.59
	Recall	0.57	0.64	0.54	0.58	0.75	0.61
	F1-Score	0.54	0.60	0.54	0.57	0.71	0.59
	Training Accuracy	0.79	0.79	0.78	0.79	0.78	0.79
	Testing Accuracy	0.56	0.63	0.54	0.57	0.75	0.61
5. Proposed model	Precision	0.7	0.56	0.78	0.82	0.74	0.72
	Recall	0.8	0.78	0.64	0.77	0.99	0.79
	F1-Score	0.67	0.72	0.68	0.75	0.78	0.72
	Training Accuracy	0.94	0.94	0.94	0.94	0.95	0.94
	Testing Accuracy	0.78	0.88	0.75	0.80	0.71	0.78

Table 8 compares the results of the precision, recall, F1-score, training, and testing accuracy of the proposed model with other traditional machine learning algorithms. After exploring different machine learning algorithms, Our proposed model using a sequential neural network and multi-target regression model with a TF-IDF vectorizer got high scores in precision, recall, F1-score as compare to other models. The average precision score of our proposed model is 72%, recall is 79%, F1-score is 72%, the training accuracy is 94%, and testing accuracy is 78%. It has the highest accuracy among all other classification models.

7 CONCLUSION AND FUTURE WORK

In the proposed analysis, an efficient technique is developed to predict the personality traits of social media users as social media is one of the most promising platforms to predict human personality. Social media is a rich platform where people can easily express their feelings without any hesitation. As discussed previously, we have designed a multi-label personality model to predict Twitter users' personality traits. The algorithm of myPersonality data is used to train the model and a multi-target regression and deep sequential neural network model are used for the prediction. myPersonality dataset is a labeled data. To vectorize the data TF-IDF is used. The proposed model has achieved more reliable results as compared to other vectorizers and machine learning algorithms. Its accuracy is 94% for the training data and 78% for the testing data, According to the prediction of the proposed model, most of the tweets are related to agreeableness and openness traits of the personality. People are more agreeable and open in their tweets. Imran Khan has the highest scores in openness and Bill Gates has the highest traits in agreeableness. The people who are related to openness use terms like enjoy, happy, congratulations, great, pretty, awesome, etc, The conscious people use the terms like thoughts, feeling, dying, failure, terrified, etc, The extrovert people use terms like party, love, sing, song, etc, and the agreeable people use terms such as healthy, kind listen, believe, etc, Finally, the neurotic people use terms like regret, loosing, confuse, upset, mess, etc. It is obvious from the literature review that there has been tremendous work in the field of personality assessment but it still needs more research as personality prediction is a broad domain. In our research, We are predicting human personality through text analysis and the work can be extended by using images, videos, audio content that the social media users share on their accounts. Apart from Twitter, there are several social media sites like Instagram, Youtube, LinkedInthat can be used to explore personalities for the proposed technique. Social media users belong to different cultures and religions so they speak different languages, therefore the language barrier is one of the ultimate problems while predicting a user's personality.

8 ACKNOWLEDGMENT

First and Foremost all worship, honor, and praise be to Almighty Allah, who is the greatest of all and we all depend on Allah for supervision and sustenance. Almighty Allah gave the strength, courage, and potential to do this research. The authors are grateful to the anonymous researchers for reviewing our work and giving valuable ideas.

9 REFERENCES

[1] Caprara, G. V., Barbaranelli, C., Borgogni, L., &Perugini, M. (1993). The “Big Five Questionnaire”: A new questionnaire to assess the five factor model. Personality and individual Differences, 15(3), 281-288.

[2] McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. Journal of personality and social psychology, 52(1), 81.

[3] Sitaraman, G. (2014). Inferring big 5 personality from online social networks (Doctoral dissertation).

[4] Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. Cerebral Cortex, 24(8), 1979-1987.

[5] Ahmad, N., & Siddique, J. (2017). Personality assessment using Twitter tweets. Procedia computer science, 112, 1964-1973.

[6] Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. Journal of personality and social psychology, 81(3), 524.

[7] Wu, X. K., Zhao, T. F., Chen, W. N., & ZHANG, J. (2020). Toward Predicting Active Participants in Tweet Streams: A case study on Two Civil Rights Events. IEEE Transactions on Knowledge and Data Engineering.

[8] Rohit, G. V., Bharadwaj, K. R., Hemanth, R., Pruthvi, B., & Kumar, M. (2020, August). Machine intelligence based personality prediction using social profile data. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1003-1008). IEEE.

[9] Kunte, A. V., &Panicker, S. (2019, November). Using textual data for personality prediction: a machine learning approach. In 2019 4th International Conference on Information Systems and Computer Networks (ISCON) (pp. 529-533). IEEE.

[10] Carducci, G., Rizzo, G., Monti, D., Palumbo, E., &Morisio, M. (2018). Twitpersonality: Computing personality traits from tweets using word embeddings and supervised learning. Information, 9(5), 127.

[11] Kircaburun, K., & Griffiths, M. D. (2018). Instagram addiction and the Big Five of personality: The mediating role of self-liking. Journal of behavioral addictions, 7(1), 158-170.

[12] Varshney, A., &Puri, S. (2017, January). A survey on human personality identification on the basis of handwriting using ANN. In 2017 international conference on inventive systems and control (ICISC) (pp. 1-6). IEEE.

[13] Gupta, N., Waykos, R. K., Narayanan, R., & Chaudhari, A. (2017). Introduction to machine prediction of personality from Facebook profiles. Int. J. Emerg. Technol. Adv. Eng, 66-70.

[14] Yakasai, A. M., & Jan, M. T. (2015). The impact of big five personality traits on salespeople's performance: Exploring the moderating role of culture. Kuwait Chapter of the Arabian Journal of Business and Management Review, 4(5), 11.

[15] Dehghanan, H., & Rezaei, M. (2014). A study on effect of big five personality traits on emotional intelligence. Management Science Letters, 4(6), 1279-1284.
[16] Alam, F., Stepanov, E. A., &Riccardi, G. (2013, June). Personality traits recognition on social network-facebook. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 7, No. 1).
[17] Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011, October). Predicting personality from twitter. In 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing (pp. 149-156). IEEE.
[18] Stillwell, D. J., & Kosinski, M. (2004). myPersonality project: Example of successful utilization of online social networks for large-scale social research. American Psychologist, 59(2), 93-104.