Check for updates

SAGE Open
July-September 2021: 1–15
© The Author(s) 2021
DOI: 10.1177/21582440211032156
journals.sagepub.com/home/sgo
SAGE

# A Neural Network Approach for Predicting Personality From Facebook Data

**Seren Başaran[1] and Obinna H. Ejimogu[1]**

## Abstract

Everyday, social media usage particularly Facebook usage are growing exponentially. Simply, inspecting Facebook usage provides meaningful information concerning users' daily interactions and hence about their personality traits. Numerous studies have been done to harness such streams of Facebook data to obtain accurate prediction of human behavior, social interactions, and personality. The aim of this study is to build a neural network–based predictive model that uses Facebook user's data and activity to predict the Big 5 personalities. This study combines the inference features highlighted in three different relevant studies which are; number of likes, events, groups, tags, updates, network size, relationship status, age, and gender. The study was conducted on 7,438 unique Facebook participants obtained from the myPersonality database. The findings of this study showed how much a person's personality can be predicted only by analyzing their Facebook activity. The proposed artificial neural network model was able to correctly classify an individual's personality at an 85% prediction accuracy.

## Keywords

artificial neural network, deep learning, Facebook, multi-label classification, personality prediction

## Introduction

### *Background*

Over the last two decades, social media has rapidly become an integral part of our lives. Our dependence on social media tools and staying connected through them have gain rapid momentum particularly during present covid-19 pandemic lockdown measures in many areas from education to health and entertainment. Social media sites like Facebook, Instagram, Twitter, You Tube, WhatsApp, and Research Gate are based on the aim of providing users a milieu to communicate and share users' knowledge, experiences, opinions, and various moments of their lives. As a result, a voluminous amount of data is constantly being emerged from these social media sites based upon user generated interactions.

Such massive information has caught the attention of many researchers who thrive to predict human behavior from those interaction data. In particular, today Facebook has more than 2.8 billion active subscribers (Statista, 2020). Accommodating such immense and diverse personal information, mining and harnessing Facebook data has been the attention of many academic and business endeavors (Wilson et al., 2012). A lot of personal information is constantly being uploaded on Facebook. The user interaction data can be used to extract meaningful information regarding real-life behaviors of users. Analyzing such acquired data can also be used to discover more about users' future actions.

With this in mind, it is plausible to ask what more can be derived from such data, that is why personality prediction has become an important aspect of social media research and big data analytics. Particularly, some research studies focused on predicting personality from various social media tools that have become a popular academic endeavor. Among many studies in the extant literature, some research remarked the existence of important association among personality and various Facebook activities such as likes, tags, status updates, locations, friends, events, and posts (Amichai-Hamburger & Viniztky, 2010; Kalghatgi et al., 2015; Laleh & Shahram, 2017; Tandera et al., 2017; Tareaf et al., 2019; Xue et al., 2018; Zhu, 2020). Some examples regarding the importance of predicting personality from Facebook user generated data could be as follows.

Being able to use Facebook data to understand the personality of the users, companies can harness this information to expand their businesses and reach their target market. People who have high tendency to commit crimes can be easily predicted using Facebook data and people can also know the personality of people before going into any relationship with them.

[1]Near East University, Lefkoşa, Cyprus

**Corresponding Author:**
Seren Başaran, Computer Information Systems, Near East University, via Mersin 10 Turkey, Lefkoşa 98010, Cyprus.
Email: seren.basaran@neu.edu.tr

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Different techniques have been applied so far in literature and numerous studies have shown that there is a certain linkage between users and their Facebook profile. This association can be harnessed and be applied into different areas such as targeted marketing, psychology and more (Golbeck et al., 2011).

Using Facebook data to determine a person's personality based upon the Big 5 personality model can be classified as a "multi-label classification" (MLC) problem, in the sense that an individual can possess more than one personality trait. In the Big 5 model, personality of a person differs in terms of openness, conscientiousness, extraversion, agreeableness, and neuroticism (OCEAN; Costa & McCrae, 1992). Each of these five personality traits all corresponds to a classifier. An MLC problem is defined as where more than one target label is attached to each instance. This method is often applied to tasks such as text categorization, medical diagnosis, music categorization, and semantic scene classification (Tsoumakas & Ioannis, 2006). An individual can be categorized under more than one personality label in a MLC problem.

Different techniques have been proposed to solve problems such as those, some of which are Multi label K Nearest Neighbors (ML-KNN; M.-L. Zhang & Zhou, 2007), Artificial Neural Network (ANN), Naïve Bayes, support vector machine (SVM) Decision Trees, and Logistic Regression (Hall, 2017).

ANN is a type of multi-dimensional regression analysis model, which makes it in various ways better than conventional regression models. The inspiration behind the development of ANN is stemmed on developing an intelligent system that can perform task intelligently like that of a human brain (Devi et al., 2012). Regardless of how complex a system might be, ANN can accurately perform prediction. That is why a lot of researchers use ANN for prediction problems especially in cases where the problem is too complicated to be expressed in a mathematical formula and also in a case where the input/output data are available (Bataineh et al., 2016).

This study aims to use ANN to predict personality with the dataset derived from Facebook. The dataset retrieved from myPersonality database (Kosinski et al., 2015) consists of more than 3 million Facebook users.

Some studies use linguistic behavior of a person from a person's status update to predict personality (Tandera et al., 2017), but this research sought to predict personality by analyzing and utilizing the relationship between a user's personality and their Facebook activities.

The personality of an individual is stable through time and situation (Espinosa & Rodríguez, 2004), meaning personality of an individual does not change online or offline, an individual that is sociable offline will be sociable online. Therefore, the Facebook profile of an individual can reflect actual personality (Back et al., 2010).

There are some studies in literature that predicts Big 5 personality utilizing features such as linguistic which is retrieved from written text or speech text (Mohammad & Kiritchenko, 2013). However, the topic of predicting personality on social media has become a popular one.

The back propagation (BP) algorithm for neural network was typically used in ANN studies but since the dataset to be analyzed involves a multi-label classification problem, some important characteristics of multi-label learning are not captured with the basic BP algorithm, which does not consider correlations of different labels. A modified BP algorithm is better suited for MLC problems used in this study.

There are significant relationships between an individual's personality and their Facebook activity, this is to say that based on a person's Facebook activity one can get clues to a person's personality (Sumner et al., 2011). This study investigates to show whether the similarities between an individual's personality and their Facebook activity can be used to better predict personality more successfully.

This study contributes to an expanding literature on inferring personality with social media by using back propagation feed forward algorithm to analyze the Facebook activity data to see if better prediction results can be achieved. Upon the completion of this research, there was no knowledge of any literature that uses neural network strictly together with Facebook activity without looking at post and text to predict personality. Also present studies use a small-scale data set for analysis which might impede the reliability of their results.

This research practically contributes to the field by investigating the linkage between a user's Facebook activity and their personality by using a neural network predictive model to analyze information from the users Facebook activity data which will help us to understand the extent of such relationship and to know if this can help better predict a user's personality more accurately. A model that can accurately predict personality may help adaptive applications adjust to user behavior accordingly. Many examples of ANN driven personalized advertising, customized education, and viewing posts on Facebook are already available.

## Related Research

This study involves three important aspects: the study is a multi-label classification problem which uses artificial neural network to predict Big 5 personality traits of users from Facebook data. Therefore, the literature was divided into the following sections accordingly as Big 5 personality, multilabel classification, and artificial neural network and its use in prediction.

### *Big 5 Personality*

There are five major characteristics that define human personality known as "Big 5," this is a well experimented and scrutinized structured for individual personality used by researchers recently (Goldberg, 1992). The Big 5 personality traits are classified as openness, conscientiousness, extroversion, agreeableness, and neuroticism as shown in

**Table 1.** Big 5 Personality Traits (John & Srivastava, 1999).

| Openness | Conscientious | Extroversion | Agreeableness | Neuroticism |
|---|---|---|---|---|
| Imaginative, Wide interest, Curious, Intelligent, Artistic, Unconventional | Organized, Disciplined, Planner, Goal oriented, not impulsive | Energetic, Forceful, Adventurous, Enthusiastic | Sympathetic, Straight forward, Compliance, Generous | Anxious, Tense, Worried, irritable, impulsive, shy |

Table 1. Over the years, this Big 5 model has become a standard for personality due to the fact that it came out of prior test on personality, and the test also showed that the models validity was not altered by languages or variation in method analysis (McCrae & John, 1992), therefore resulting in its acceptance. When dealing with the Big 5 personality model, each individual can highly exhibit some of these traits together therefore meaning that the personality traits are not contrasting to each other. A person can exhibit high symptoms of Agreeableness, Openness, while exhibiting little symptoms of Neuroticism. For instance, a person who is an extrovert in real life always tends to post a lot about their activities and share their experiences while a person who is neurotic often tends to be less active and have less tags due to their shy nature.

## Multi-Label Classification

In machine learning, multi-label classification (MLC) is a form of classification problem but varies differently from other classification problems, in the sense that each sample can have several labels (Tsoumakas & Ioannis, 2006).

This varies from other classification problem that can have just one label and never two (i.e., an object can either be classified as dog or cat but never both) and this is known as Multi-Class Classification. In MLC, samples are attempted to be classified in more than one label (i.e., a person can be both labeled as openness and agreeableness; Tsoumakas & Ioannis, 2006).

Algorithm Adaptation uses certain algorithm to directly alter and classify standard classification technique to perform MLC. This schema treats MLC as a single integrated problem without requiring problem transformation. Some examples of machine learning methods that have adapted this approach in handling MLC are ANN, boosting, decision trees, and KNN (Hall, 2017). For the case of the Big 5 personality traits which are independent of one another, an individual can exhibit high symptoms of more than one personality trait hence making it a multi-label learning task and among other approaches, ANN approach presumed to yield better results in predicting personality accurately.

## Artificial Neural Network

ANN are designed to act like the biological nervous systems work in interacting with objects of the real world, they are a large parallel interconnected networks made up of nodes and each node is referred to as neurons (M.-L. Zhang & Zhou, 2006).

ANN has the ability to learn, to adapt by modifying its internal structure depending on the data that passes through it. It is one of the most successful learning methods and has performed so well in classification (J. Zhang, 2016). ANN provides variations of techniques to learn from examples and performs very well in pattern recognition. At the moment various types of neural networks exist as examples: mapping networks, radial basis function networks, adaptive resonance theory models and of course multi-layer feed forward neural networks (Kalghatgi et al., 2015).

Numerous studies have been carried out in the past using ANN as a tool, only the studies conducted using ANN for prediction were examined, then after that studies carried out in the area of ANN in prediction for multi-label classification problems were examined and then finally studies relating to ANN in personality prediction.

## Using ANN for Prediction

Different models and methods have been proposed for prediction of various outcomes. In 2010, ANN was used as a tool to predict team performance by analyzing individual past achievements and history (Hedberg et al., 2010). The aim of the study was to provide a means by which employers can analyze prospective team member's track record to understand the effect of that individual in the team. After analysis, training, testing, and evaluation, the model achieved 73.4% prediction accuracy. With this level of accuracy, the study claims that this ANN approach can be applied in other organizational levels including recruitment which also forms a basis for predicting personality as well.

Champa and AnandaKumar (2010) conducted a study on human behavior prediction through handwriting analysis. The study uses ANN to analyze various samples of individual handwriting by looking at the baseline, the pen pressure, and the letter "t." The study states that professional handwriting examiners can understand human personality from individual's handwriting; however, the process is costly and prone to fatigue. The baseline, the pen pressure, and the height of the t-bar in the letter "t" stem were fed into the ANN as inputs and outputs as individual personality traits. The model was run through various epochs and hidden layer and attained a maximum accuracy of 53%.

Another study by Nkoana (2011) proposes an ANN model for flood prediction and early warning, in the study various number of trained neural network architectures were evaluated using their mean percentage accuracy. The study implemented 14 neural networks using daily rainfall as the predictive variable from the period of 1995 to 2009, after examining the performance of the neural networks the Elman recurrent neural network with two hidden layers and two hidden nodes yielded a better result of 58% accuracy. The study claims that using ANN with daily rainfall can be used to predict floods. Another study by Devi et al., 2012 also proposes an ANN model for Weather prediction. The study collects data from atmospheric pressure, temperature, wind speed, wind direction, humidity and precipitation and uses it to train a three-layer ANN. The results were compared with practical working of the meteorological department and the study claims to have built a model which can successful predict weather based on the comparison results.

Another interesting study using ANN for future predictions was by Song and Kim (2014), the study feeds the Big 5 personality trait as input into the ANN model to predict individuals' future location. That study explores the connection between human mobility patterns and their personality to train the ANN to predict future locations. The study combined time information and personality as input nodes while locations as output sample training data. The researcher claim to predict human location through the help of the personality trait properly. The study recommends to use the reverse of this model in the future to use mobility pattern to predict personality.

Binh and Duy (2017) used ANN as a tool to predict student performance based on the students learning style. The study conducted an online survey with a participation of 316 undergraduate students in various courses. Using the data collected and analyzed an ANN model was built to predict students' performance based on their learning style. The ANN model managed to produce 80.63% classification accuracy, the study claims that this can method can be applied in e-learning environment adaptive models that can support learners.

Al-Shihi et al. (2018) proposed a model that can be used to predict mobile learning adoption in developing countries. The study integrates some constructs such as social learning, flexibility learning, enjoyment learning and economic learning. The study was conducted on 388 participants from major universities/colleges at Oman and ANN was used as the tool for prediction. The study suggests that proposed model can be used to predict and influence mobile learning adoption.

## Using ANN for Prediction Through Social Media

Nam et al. (2014) proposed a simpler ANN approach to handle multi-label classification in large-scale multi-label text classification. The proposed method is aimed at being an alternative and better method than the state-of-the art back propagation multi-label learning approach. In the study, the BP-MLC's pairwise ranking loss was replaced with cross entropy, and other features such as ReLU activation function was used together with AdaGrad optimizers. The study claims that this approach enables the model converge in just a few steps and the dropouts utilized helps prevent overfitting. The study evaluates the performance of the proposed model with other baseline models. The algorithm trains with a higher convergence speed due to the ReLU activation, the model also uses dropout to prevent overfitting by randomly dropping individual hidden units while by taking advantage of label space inherent correlation to minimize rank loss.

Liu and Chen (2015) proposed a multi-label approach for sentiment analysis of microblogs. The study compares 11 state-of-the art ML classification methods and uses eight metrics for evaluation. The comparison was carried out on two microblog datasets. Out of the 11 methods evaluated, some of the methods performed better than others depending on the scenario. Rakel (Random K label set) performs better with HR, while other algorithms performed better on AI. So, the different features in the results affected the results of the study but the result of the study shows that one of the dictionaries used in the study Dalian University of Technology Sentiment Dictionary performs best on multi-label classification.

Corani and Scanagatta (2016) proposed a multi-label classifier model which is based on Bayesian networks but performs slightly different from the baseline Bayesian network. The model addresses the dependencies among the class variable which is normally overlooked when devising independent classifier for each of the classes to be predicted. The model works by simultaneously predicting the class variable which is different from the baseline approach, the study results show that the performance of the proposed model out performs the independent approach when predicting multiple air pollutions.

Kee et al. (2017) proposed a neural network multi-label classification system to predict the arrival time of bus transport. The neural network is built based on the historical GPS (Global Positions System) arrival time and ensemble of neural network is used to improve the reliability of the output. The results of the study show that the proposed model is able to forecast the arrival time up to a reasonable percentage of 75%. The neural network and ensemble model was compared with other algorithms such as decision tree, Random forest, Naïve Bayes, and the model proves to be 8% better than the other algorithm. The study suggests further improvement of the model by using power transformation and some other different ensemble methods.

## Personality Prediction Through Social Media

Wald et al. (2012) proposed a form of machine learning ensemble learning called SelectRUSBoost to predict psychopathy through Twitter data. This method adds feature selection an imbalance aware ensemble to tackle high dimensionality. The study states that when ensemble learning, data sampling and feature selection in SelectRUSBoost, the model is able to hit AUC (Area under the curve) of 0.736 and this performance is only achieved when this model is used. The study states that a model such as this can be used by law enforcement in discovering psychopathic cases through their Twitter data. The study also states that though the model can be used with Twitter to predict the incidence of psychopathic situations. They are not sufficient to provoke direct actions but can be used to flag potential risk.

Farnadi et al. (2013) explored the use of machine learning (SVM, NB [Naïve Bayesian method], KNN) to infer personality just by examining Facebook status updates of various users. The study strengthens their prediction model by not just relying on one source but by including different training samples from another source (Essay corpus) helping the study show that trait can be generalized across social media platforms. The study investigates 250 users with 9,917 status updates and states that despite having a small amount of dataset the model could still outperform other baseline methods.

Another study by Kandias et al. (2013) proposes a methodology that detects users that are hostile or with a negative attitude toward the authorities, the study combines the dictionary learning-based approach and machine learning techniques (SVM, NB, LR [logistic regression]). The study analyzed information posted on the YouTube website.

Lima and de Castro (2014) in their study use a semi supervised classification approach to predict personality through twitter data. The data take a different approach from other studies, this study does not take user profile into consideration, and it does not work with single texts like in other studies but works with a group of text. The study uses the problem transformation method to transform the problem into five binary classification problems. The study used three well-established machine learning algorithm: NB, MLP (multi-layer perceptron), and SVM to train the proposed system and was applied to predict personality from tweets which resulted in an 83% prediction accuracy.

Kalghatgi et al. (2015) also investigated Big 5 personality trait prediction through analyzing tweeter data with ANN. The study explores the parallelism between an individual's linguistic information and their Big 5 personality trait and uses the tweets posted by an individual to predict personality. The study also reports that the model doesn't take user tweeter profile into consideration and implements it in java NetBeans using Hadoop framework to make predictions of multiple individuals at the same time.

Akshat (2016) investigates using CNN to predict personality from social media images, the study sort to find out if there was any relationship between the output why such relationship exist. The study results show how powerful neural network is as a tool to measuring and learning highly nonlinear mappings between input data and output data. The study uses the transformation method to transform the task into a classification task and uses a chance baseline which guesses just the highest occurring class which is used for comparison. The model was trained and validated with a split of 80, 10, and 10 for training, testing and validation.

In 2017, Tandera et al. (2017) carried out a competitive analysis of current deep learning architecture and uses accuracy results to compare performance. The study involved using the models to predict Big 5 personality trait from data retrieved from users Facebook account. The dataset used in the study were gotten from two different sources; myPersonality dataset consisting of 250 users and then 150 Facebook users data which were collected manually. The study also uses linguistic features such as LIWC with both closed and open vocabulary approach. The study reports saying the model outperforms other methods by 74.14% average accuracy, though accuracy was low with some traits, study claims this could be a result of limited dataset. The experiment results show ANN doing better than other traditional machine learning classification method.

Again in 2017, Laleh and Shahram proposes a model that uses LASSO algorithm to select the best features and predict the Big 5 personality trait from a user's Facebook data by examining Facebook likes. The study examines the likes of 92,225 users while combining with 600 weighted topics, the model also examines the task as a regression problem. The training and test data are split 75% and 25%. The cross validation method was used to validate the model. Still in 2017 is a study by Iatan which uses Fuzzy Gausian Neural Network (FGNN) to predict personality from a user's Facebook account based on the data publicly available and compares result with two other models; multiple linear regression model and multi-layer perceptron. The performance of the model was tested using normalized root mean square. The study results show how the proposed method outperforms the other two methods during training.

## Method

There is a relationship between an individual's Facebook profile and their personality; there are some predictive models that take into account the Facebook activities of users and their networks (Bachrach et al., 2012). The study attributes such as number of likes, groups, tags, and friendship networks were the features focused on. Another study (Kosinski et al., 2013) also proposed a predictive model that just focuses on the demographic information retrieved from the users' Facebook profile and demographics such as age, gender, and relationship status. All these previous works show tight relationship between the users Facebook profile regarding their usage pattern and their demographics. Based on this, in this study, a different dataset is used and a combination of two different predictive models from previous work is used to formulate the predictive model for this study. The model used was built on the findings of earlier three studies.

The model is a combination of the features highlighted by Bachrach et al. (2012) and Sumner et al. (2011) with the

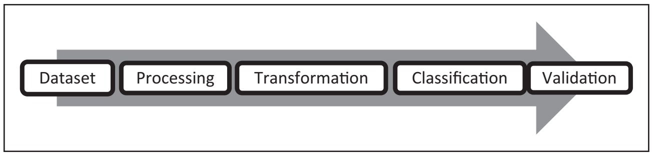


**Figure 1.** The classification process of BPNN model.

features highlighted by Kosinski et al. (2013) in their studies which are number of likes (total number of like button pressed), number of status updates(the total number of Facebook status updates occurred), number of events(the total number of Facebook events joined or created), number of groups (number of Facebook groups joined), number of tags (total number of Facebook post tags used), network size(total number of friends), relationship status(the status showing the relationship), age(age of the user), and gender(gender of the user). The features with potential high influence on personality prediction were chosen for this study.

There are some steps that need to be taken when creating an algorithmic model. The first step and most crucial is the pre-processing of the data. This step is what prepares the data for the task to be carried out. Feeding data that have not been properly processed can greatly hamper the results of the model and can throw off the model completely. Before classifying or feeding the data into the ANN network first processing must be done. The next step is the actual processing itself which involves transforming the data received and then finally feeding it into the ANN for classification. Figure 1 shows the flow diagram of the model used in this study.

The Dataset used for this study was obtained from the database provided by the myPersonality project (Kosinski et al., 2015) which consists of Facebook data of more than 4 million participants with given personality label which are based on the Big 5 personality model. The myPersonality project was initiated by David Stillwell and Michal Kosinski. It is a Facebook application that collects users Facebook information from their Facebook profile while taking privacy issues into consideration and also allows them take psychometric tests which calculates things such as satisfaction with life and Big 5 personality. The data retrieved from the application were processed, analyzed, and then used to create the datasets. The data contain information of user's demographics, activities, and friendship network size. During this study, the following datasets were downloaded.

- *Big 5* model personality score: These data contained Big 5 personality test scores taken by 3,137,694 Facebook users. It contained scores for the main Big 5 traits Openness, Neuroticism, Agreeableness, Conscientiousness, and Extraversion which results were scaled from 0 to 5.
- *Facebook activity*: These data contained a summary of the activities (tagging, posting, joining groups etc.) of 1,674,259 Facebook users. This table contained the number of likes, tags, updates, events, groups, and friends.
- *Demographics*: In this data, the basic attributes exist such as birthday, age, gender, relationship status, interest, time zone, and network size of 4,282,857 unique Facebook users.

The dataset were downloaded from the myPersonality database and needed to be merged together into one file, Microsoft SQL was used to merge the various database by their unique user id. The script used to merge the various database can be found in the Appendix section.

The different database did not contain equal amount of participants, so to merge them, only common participant that could be found in all three Databases were merged, others that could not be found in the other databases were dropped. After merging the data, the dataset was left with 1,337,313 rows of unique participants with unique user IDs.

After various database files were successfully merged, missing values were considered. Missing data can greatly affect the results of a research because it could lead to biases, affects the findings, generalizability, and result to a great loss of information (Dong & Peng, 2013). To ensure no missing values in the data, missing value analysis was done using IBM SPSS, and it was observed that the merged file had a lot of missing values and before the database can be used for the study, missing values must be addressed. Some of the tables in the dataset had missing values above 50%, so two methods had to be used to deal with the missing data, listwise deletion, and replacing using series mean (Humphries, 2013). Another script was written in MSSQL to handle missing values. Starting from the column with the highest missing, the script compares the values in the column with other columns and deletes that row if a missing value is found; this step was repeated across the columns until missing value was below 10%, eventually reducing the data to 7,438 participants. After the missing values were reduced to 10% or less, the replace missing value option using series mean in SPSS was used to replace the remaining of the missing values.

Now dataset only comprises participants with no missing data, the data could now be further processed to be used in the neural network model (See Figure A1).

Out of the 7,438, 3,013(40.5%) were male and 4,425(59.5%) were female. The majority (57.4%) of the participants were in the age group of 18 to 25 years, followed by those (28.1%) within the age group of 26–40 years, then the 7.4% between 18 years and below, the 6.4% within 40–60 years, and also 0.7% those with 60 years and above. The Dataset also shows Big 5 personality traits of the participants; 96% exhibited openness traits and 4% did not, 57% exhibit traits of neuroticism and 43% did not, 91% exhibited agreeableness traits and 9% did not, 87% exhibited conscientiousness traits but 13% did not and finally 88% exhibited extrovert traits and 12% did not. Table 2 represents the actual

**Table 2.** Big 5 Personality Distribution.

| Value | OPE | NEU | AGR | CON | EXT |
|---|---|---|---|---|---|
| Yes | 7,181 | 4,209 | 6,789 | 6,502 | 6,567 |
| No | 257 | 3,229 | 649 | 936 | 871 |

number of participants who displayed the Big 5 personality traits after preprocessing the dataset.

When feeding the data into the neural network, the data to be fed should be in tensors of floating point data (Chollet, 2017). The data also must not take widely different ranges because it could affect training.

The values in the dataset are all in different scales and in order for the neural network to properly work with the data, the data had to be rescaled to become uniform. Since the neural network is to be built using Python, the rescaling of the data was done using Python also. In Python, data processing and neural network has been simplified with the help of TensorFlow by Google, every operations was carried out with TensorFlow as the backend.

Steps for rescaling were given in the following:

1. The data were loaded using Panda's module in TensorFlow, after the data were loaded, the input values and output values were defined. Input variables were 9 (number of likes, relationship status, number of status updates, number of events, gender, age, network size, number of groups, number of tags and output variables were 5 personality traits(Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism).
2. To normalize the data, some of the columns (relationship status, gender) had to be rescaled because they are nominal data. Using "One hot encoder " the column was split into ten and scaled into binary, which increased the input variable to 18.
3. The remaining of the data were normalized using the MinMax Scale (0–1) in python. This was chosen because it helps in feed forward back propagation during gradient descent calculation.
4. Since our problem is a multi-label classification problem the output data (y) was further rescaled into binary of 0 and 1 with the binaries function provided by scikit in Python. Values greater than 0.5 were classified as 1, while values less than 0.5 were classified as 0.

To ensure this is not the case, the best approach will be to normalize the data by transforming them into vectors of −1 to 1 or 0 to 1. Some of the data also might be scaled from 1 to 10. If there are 10 variables, it takes the first variable and turns it into 1 and then turns the rest into 0's, it then takes the second integer, turns it into 1 and then turns the rest into 0's it continues this process for all the remaining integer. After the dataset has been successfully normalized and transformed it can be sent further for classification which outputs will return either 0 or 1 (See Figure A2).

Various methods have been used for classification in the past but in this study a multi-label back propagation neural network technique was utilized for classification. ANN has been widely adopted into multi-label classification. The configuration and parameter in an ANN model needs to be selected efficiently so as to ensure appropriate generalization and efficient learning. What the model does is that through a feed forward process, updates itself by the back propagation update method and uses the supervised topology to enhance the model. This method is a multi-purpose learning algorithm, very effective and produces great results but it also costly in terms of learning requirement. The hidden layer is the layer between the first and the last output layers. The data are taken from each of the input neurons through the synapses and multiplies it with a set of random weights. The summed weighted inputs are then passed through an activation function to the output layer. In this study, two activation functions were used ReLU and sigmoid activation function. The first to be used is the ReLU activation function. ReLU is the most used activation function in the world today when sending signals from the first layer to the next layer before the output layer (Sharma, 2017). Also since this is a multi-label classification problem and each label prediction probability needs to be predicted independently of the other class probabilities, sigmoid activate function is used as the second activation function

$$\left(\mathbf{A}\left(\mathbf{x}\right)=\mathbf{max}\left(0,1\right)\right)$$

$$\left(\mathbf{A}=\frac{1}{1-\mathbf{e}^{-\mathbf{x}}}\right)$$

An ANN task with multiple possible label samples that are not mutually exclusive meaning that a sample can have multiple label and not restricted to just one label, this is known as a multi-label classification problem. This problem is well tackled in ANN with a framework known as keras. In this study, we have a problem with five different labels (openness, agreeableness, neuroticism, conscientiousness and extraversion); therefore, this study has *n* samples

$$\boldsymbol{X}=\left\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{\boldsymbol{n}}\right\}$$

and *n* number of labels

$$\boldsymbol{y}=\left\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{\boldsymbol{n}}\right\}$$

with *y_i* ∈ {1,2,3,4,5} and *P*(*c_j*—*x_i*) for the prediction probability. The next thing is to build a simple ANN with five output nodes and one output for each class. Designing the input and hidden layers is quiet straight forward but designing the output layer for a multi-label and choosing what kind of layer it will be is quite important. Usually the softmax layer is the choice for multi-classification problem

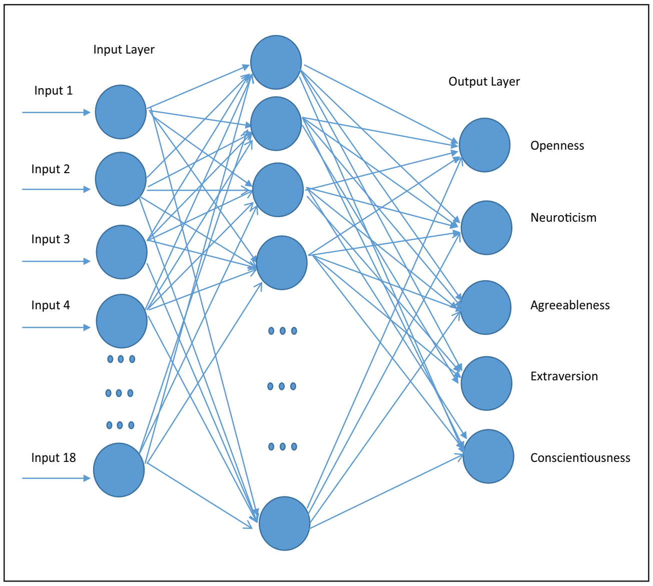


**Figure 2.** Neural network model.

but this is not really the best choice for a multi-label problem (Sterbak, 2017). In softmax when increasing score for one labels all others are lowered (probability distribution) which is not a problem when predicting a single label per sample but in a multiple label prediction this is not good.

What is needed is to decompose the multi-label classification task, for this, a sigmoid output layer is need consisting of a sigmoid activation function and binary_crossentropy loss function. The labels will be improved individually and each label is independent of the other labels probability. There are 18 inputs and 5 outputs as shown by Figure 2.

$$P\left(c_j \mid x_i\right) = \frac{1}{1 + exp\left(-z_k\right)}$$

The study employed the use of the Keras API developed my Google which runs on TensorFlow to build the ANN model. This API allows models to be built easily and provides easy manipulation of data for better learning (Maxwell et al., 2017). Due to the flexibility of TensorFlow, developers can easily experiment on various optimization techniques and algorithms, which helps in simplifying implementation. This model was built using Python. Python has become one of the most popular languages for data science (Chollet, 2017), it is a language that is very comprehendible by a wide range of people and it is easy to read syntax.

In classification techniques, the dataset are separated into training, testing and validation data so as to be able to determine and monitor the accuracy of the model. The training data will consist of 60% of the data; these data are examples by which the network can use to learn patterns and variations to make its decisions these data will be fed to the training model repeatedly. While the training is going on, another 20% of the data are used to validate the quality of training, if this looks good the remaining 20% of data, which has not been exposed to the network, is used to test the accuracy of the model.

There are two common approaches used to evaluate the performance of a classification model that are K-fold and leave one out validation (Wong, 2015). In this study, the K-fold cross validation was used for this study to conduct the evaluation of the model.

## Results

To carry out the data processing and modeling, python was used due to the vast amount of machine learning libraries available in the python language and the incredible data visualization that can be carried out also in the python environment.

In the building of the neural network it is important to identify the number of hidden layers and hidden neurons to be used in the network. Depending on the data it is safe to start with a few hidden layers preferably one fully connected hidden layer (Chollet, 2017).

The model was setup with a fully connected hidden layer between the input layer and the output layer. The ReLU was used to act as the non-saturating activation function between the first layer and the hidden layer while the sigmoid activation function was used as the final output activation function. When deciding for the number of hidden neurons it could be somewhat complicated to select the best number of hidden neurons perfect for the task without examining several models. Inadequate or too much hidden neurons could lead to over fitting or underfitting.

In neural network, a lot of try error is usually done to ascertain the best parameters for the network. There is some rule of thumbs when deciding the number of hidden neurons (Heaton, 2008).

- It should be less than two times the input layer size
- Two-third of the sum of input and output neurons

In this study, different amount of hidden neurons were tried and then increased to ascertain the neurons with best performance.

First, the data were imported using the panda's data frame because of the massive functionality the panda's data frame gives to work on data. After data have been successfully imported, the next step is to create a matrix of the features and target variable, and this enables the network to identify the input and the output file, 9 input variables and 5 output variables. As discussed in the previous section, features such as relationship status were coded from 1 to 10 but leaving it this way would mean 10 is higher than 1, meaning single is higher than divorced which is not the case. To handle these, dummy variables were created using a function from the SciKitLearn library in python known as OneHotEncoder (Pedregosa et al., 2011). After one hot

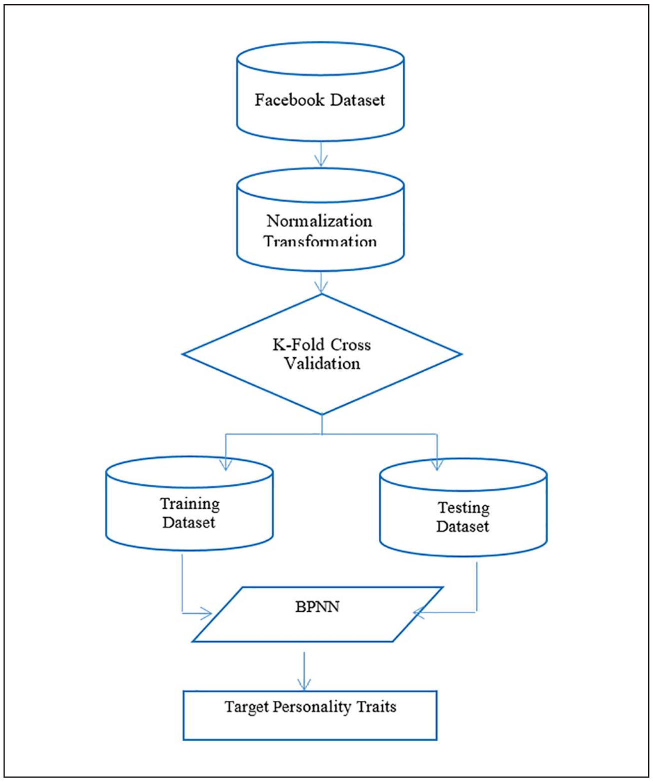


**Figure 3.** The study flowchart.

encoding, the transformation and normalization were also done using the SciKit-Learn library. The dataset is transformed into vectorized form from 0 to 1 so as to enable the network better understand the data for classification. The output target dataset was also transformed using 0.5 as its threshold. The SciKit-Learn is also used to split the data set into training and test samples.

The next step is the training and testing. This was broken down into two different schemes: the first scheme involved manually splitting the dataset and the second scheme involved using the K-Fold cross validation. The study's procedures were represented in Figure 3.

### Training

In the first scheme, which was split manually, the first part of the first scheme was split into 75% for training and 25% for testing, while the second part was split 67% for training and 33% for testing which the model has not seen before. The second scheme, however, was done using the K-fold cross validation, this was also broken into parts which was K-10 Cross validation and K-5 Cross validation

In the final process of the BPNN model, the input neuron was made up of 18 input neurons to accommodate for the 18 features. For the five personality classes; openness, agreeableness, neuroticism, extraversion and conscientiousness; the output layer was made up of five output neurons. For the sake of optimization, different parameters were set, such as learning rate, hidden neurons, and splits; the results were compared with each of them yielding different results. The maximum epochs were set to 1,000 but the network was told to end training if network loss performance keeps declining and doesn't improve after 10 epochs. All the training and computation was carried out using a 3.30 GHz PC with 8GB RAM, Intel core i5 and windows 10 OS. Table A1 shows the training parameters for the different neural networks.

### Testing

The model performs quite well with the test data and shows good generalizability. This is to say that the model gives compelling generalization abilities when presented with new users Facebook Activity Data. Table 2 shows the prediction accuracy results and the hamming loss for all the trained networks. The hamming loss is part of the metrics used for evaluation in this study. It is used to compute accuracy through the equilibrium contrast between the target data and predicted data. The hamming loss is the fraction of labels that are incorrectly predicted. The best hamming loss for the model is 14.96%, which implies that in more than 85% of the time, the model can correctly classify an individual based on his/her Facebook activity. The Table A2 represents scheme 1 and scheme 2 testing results as the prediction accuracy results and the hamming loss for all the trained networks. The learning curve showing the convergence of the network plots the loss against the epochs and the accuracy against the epochs is shown in Figure A3 to A6 representing Scheme 1 Test 1, Scheme 1 Test 2, Scheme 2 Test 1 and Scheme 2 Test 2, respectively.

## Discussion

Different studies have been carried out on the subject of using various means to infer personality. While some studies have used real-life activities pertaining to locations and speech to infer personality, a study by Golbeck et al. (2011) took a different approach and focuses on using social media to infer personality and discovered a tight relationship between a user's profile and personality. Some other studies by Bachrach et al. (2012) and Kosinski et al. (2013) shows some major inferences between demographic attributes and activities to a person's personality.

Above studies provided background motivations to carry out this study. The influence of social media and especially Facebook in the society is rapidly on the increase especially during the times of covid-19 pandemic. Facebook has become an integral part of our lives, businesses, government and many more. It is important to understand what extent of inference Facebook has to a person's personality so that this can be used to better improve and safe guard lives and lead to better society. In this Study 2, inference models from three different studies were combine, this being the features highlighted by Bachrach et al. (2012) and Sumner et al.

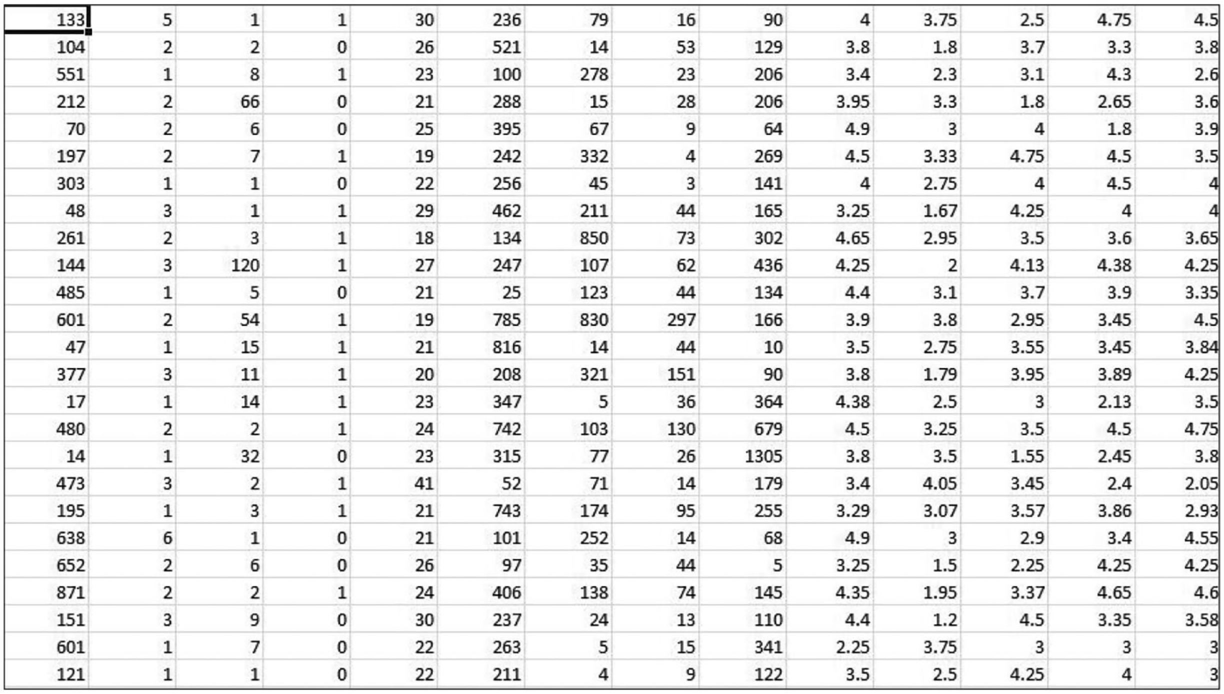
(2011) with the features highlighted by Kosinski et al. (2013). The developed framework shows its capacity to predict the Big 5 personality traits; Openness. Agreeableness, Conscientiousness, Extraversion and Neuroticism from a user's Facebook data. As shown in this study, this ANN model shows encouraging results, in that in 85% of the time the network will correctly classify a Facebook user based on just their activities. Upon trying different methods the best classification result derived from the model had a prediction accuracy of 85.04% although their differences were not so distinct. Comparing the results of this study with some other studies such Tandera et al. (2017) and Lima and de Castro (2014) networks, this model performs better. In the study by Tandera et al. (2017) which used linguistic features on Facebook data to infer personality was able to hit 70% accuracy on their neural network model, while in the study by Lima and de Castro (2014), a semi supervised learning approach was used and the outputs where broken down and analyzed separately into five different binary outputs, this method gave a 75% prediction accuracy. In this study, the proposed models analyzing just Facebook activities and demographics alone was able to perform better with a prediction accuracy of 85%. This shows how the ANN model can be used to learn accurately and faster during the training phase with if given more data; however, the generalizability is weakened in different scenarios which maybe a result of the data or the parameters used during the training. Given suitable pre-processing and adequate amount of dataset, this present study evinces the viability of ANN models for personality classification and it also shows the usefulness.

## Conclusion

The purpose of this study was to explore the performance of ANN in classifying and predicting the big five personality based on the data derived from a user's Facebook data. This study proffers an apt system classification model for Big 5 personality prediction that could accurately infer an individual's personality based on only their Facebook data with a prediction accuracy of 85.04%. The observations showed that ANN with proper parameter tuning could perform well accuracy on complex multi-label task such as personality classification when trained and tested with new data. With the rapid growth in demand among various companies in better understanding their clients, this has increased the demand for online tools that can help better under the personality of the consumers.

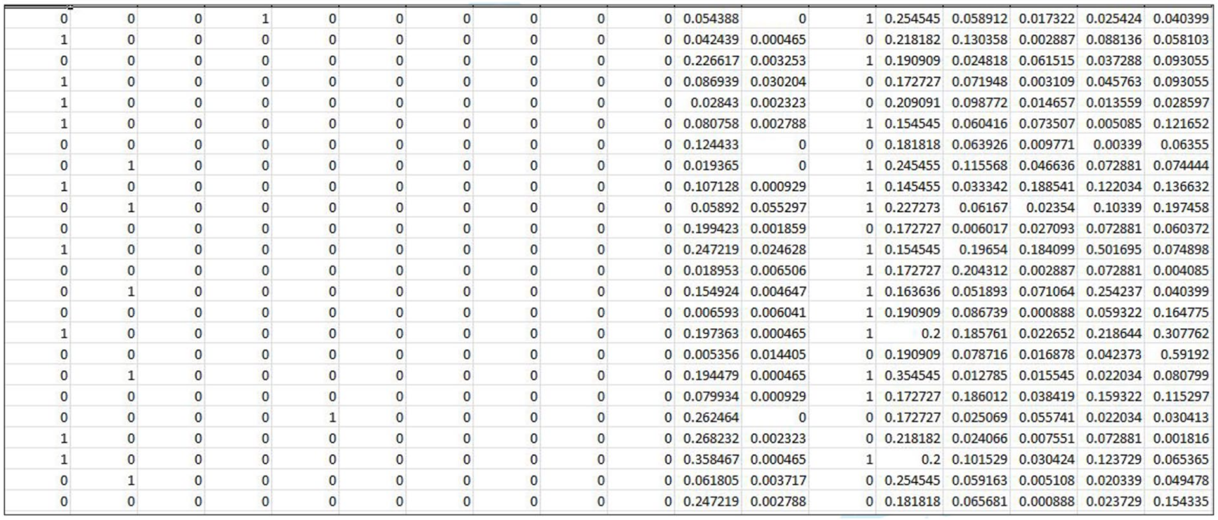
One of the limitations of this study is that a huge amount of data were lost during data pre-processing but more data can be added to the model improve training phase. To improve the model training quality, there is a need for more data, so much data can be gotten from an individual's social media account. Another limitation to this study is that accuracy was not verified with other methods such as partial least squares and other machine learning methods. This similar study should be carried out on the same participants other accounts so as to better compare results and improve prediction. Finally, more studies should be carried out in this area of utilizing neural network to better understand and predict personality so as to understand ways to make people's lives better. With the prediction accuracy improved more, this model can be implemented in Facebook, users will no longer need to fill long personal forms to be able to determine their personality type, the personality type can be determined on Facebook just from user's activities without having to fill any forms. Users can be able to make results public and share on their wall. In business, based on the requirement of company, organizations can be able to predict the personality of workers to see how they can better improve their service. Advertisers can better know how to target their audience, for instance, advertisers can target people with openness personality when advertising new products or target people with neurotic personality when advertising security products. The data that can be retrieved from Facebook data are very rich, further studies can be carried out in combining ANN with Big 5 personality traits to analyze Facebook data to predict depression and suicidal tendencies. Future work will be geared toward improving the accuracy of the model by collecting more data and cross validating the data with other social media platforms asides Facebook.

## Appendix



| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 133 | 5 | 1 | 1 | 30 | 236 | 79 | 16 | 90 | 4 | 3.75 | 2.5 | 4.75 | 4.5 |
| 104 | 2 | 2 | 0 | 26 | 521 | 14 | 53 | 129 | 3.8 | 1.8 | 3.7 | 3.3 | 3.8 |
| 551 | 1 | 8 | 1 | 23 | 100 | 278 | 23 | 206 | 3.4 | 2.3 | 3.1 | 4.3 | 2.6 |
| 212 | 2 | 66 | 0 | 21 | 288 | 15 | 28 | 206 | 3.95 | 3.3 | 1.8 | 2.65 | 3.6 |
| 70 | 2 | 6 | 0 | 25 | 395 | 67 | 9 | 64 | 4.9 | 3 | 4 | 1.8 | 3.9 |
| 197 | 2 | 7 | 1 | 19 | 242 | 332 | 4 | 269 | 4.5 | 3.33 | 4.75 | 4.5 | 3.5 |
| 303 | 1 | 1 | 0 | 22 | 256 | 45 | 3 | 141 | 4 | 2.75 | 4 | 4.5 | 4 |
| 48 | 3 | 1 | 1 | 29 | 462 | 211 | 44 | 165 | 3.25 | 1.67 | 4.25 | 4 | 4 |
| 261 | 2 | 3 | 1 | 18 | 134 | 850 | 73 | 302 | 4.65 | 2.95 | 3.5 | 3.6 | 3.65 |
| 144 | 3 | 120 | 1 | 27 | 247 | 107 | 62 | 436 | 4.25 | 2 | 4.13 | 4.38 | 4.25 |
| 485 | 1 | 5 | 0 | 21 | 25 | 123 | 44 | 134 | 4.4 | 3.1 | 3.7 | 3.9 | 3.35 |
| 601 | 2 | 54 | 1 | 19 | 785 | 830 | 297 | 166 | 3.9 | 3.8 | 2.95 | 3.45 | 4.5 |
| 47 | 1 | 15 | 1 | 21 | 816 | 14 | 44 | 10 | 3.5 | 2.75 | 3.55 | 3.45 | 3.84 |
| 377 | 3 | 11 | 1 | 20 | 208 | 321 | 151 | 90 | 3.8 | 1.79 | 3.95 | 3.89 | 4.25 |
| 17 | 1 | 14 | 1 | 23 | 347 | 5 | 36 | 364 | 4.38 | 2.5 | 3 | 2.13 | 3.5 |
| 480 | 2 | 2 | 1 | 24 | 742 | 103 | 130 | 679 | 4.5 | 3.25 | 3.5 | 4.5 | 4.75 |
| 14 | 1 | 32 | 0 | 23 | 315 | 77 | 26 | 1305 | 3.8 | 3.5 | 1.55 | 2.45 | 3.8 |
| 473 | 3 | 2 | 1 | 41 | 52 | 71 | 14 | 179 | 3.4 | 4.05 | 3.45 | 2.4 | 2.05 |
| 195 | 1 | 3 | 1 | 21 | 743 | 174 | 95 | 255 | 3.29 | 3.07 | 3.57 | 3.86 | 2.93 |
| 638 | 6 | 1 | 0 | 21 | 101 | 252 | 14 | 68 | 4.9 | 3 | 2.9 | 3.4 | 4.55 |
| 652 | 2 | 6 | 0 | 26 | 97 | 35 | 44 | 5 | 3.25 | 1.5 | 2.25 | 4.25 | 4.25 |
| 871 | 2 | 2 | 1 | 24 | 406 | 138 | 74 | 145 | 4.35 | 1.95 | 3.37 | 4.65 | 4.6 |
| 151 | 3 | 9 | 0 | 30 | 237 | 24 | 13 | 110 | 4.4 | 1.2 | 4.5 | 3.35 | 3.58 |
| 601 | 1 | 7 | 0 | 22 | 263 | 5 | 15 | 341 | 2.25 | 3.75 | 3 | 3 | 3 |
| 121 | 1 | 1 | 0 | 22 | 211 | 4 | 9 | 122 | 3.5 | 2.5 | 4.25 | 4 | 3 |

**Figure A1.** Sample of input data before one hot encoding and data transformation.



| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.054388 | 0 | 1 | 0.254545 | 0.058912 | 0.017322 | 0.025424 | 0.040399 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.042439 | 0.000465 | 0 | 0.218182 | 0.130358 | 0.002887 | 0.088136 | 0.058103 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.226617 | 0.003253 | 1 | 0.190909 | 0.024818 | 0.061515 | 0.037288 | 0.093055 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.086939 | 0.030204 | 0 | 0.172727 | 0.071948 | 0.003109 | 0.045763 | 0.093055 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.02843 | 0.002323 | 0 | 0.209091 | 0.098772 | 0.014657 | 0.013559 | 0.028597 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.080758 | 0.002788 | 1 | 0.154545 | 0.060416 | 0.073507 | 0.005085 | 0.121652 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.124433 | 0 | 0 | 0.181818 | 0.063926 | 0.009771 | 0.00339 | 0.06355 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.019365 | 0 | 1 | 0.245455 | 0.115568 | 0.046636 | 0.072881 | 0.074444 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.107128 | 0.000929 | 1 | 0.145455 | 0.033342 | 0.188541 | 0.122034 | 0.136632 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.05892 | 0.055297 | 1 | 0.227273 | 0.06167 | 0.02354 | 0.10339 | 0.197458 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.199423 | 0.001859 | 0 | 0.172727 | 0.006017 | 0.027093 | 0.072881 | 0.060372 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.247219 | 0.024628 | 1 | 0.154545 | 0.19654 | 0.184099 | 0.501695 | 0.074898 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.018953 | 0.006506 | 1 | 0.172727 | 0.204312 | 0.002887 | 0.072881 | 0.004085 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.154924 | 0.004647 | 1 | 0.163636 | 0.051893 | 0.071064 | 0.254237 | 0.040399 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.006593 | 0.006041 | 1 | 0.190909 | 0.086739 | 0.000888 | 0.059322 | 0.164775 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.197363 | 0.000465 | 1 | 0.2 | 0.185761 | 0.022652 | 0.218644 | 0.307762 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.005356 | 0.014405 | 0 | 0.190909 | 0.078716 | 0.016878 | 0.042373 | 0.59192 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.194479 | 0.000465 | 1 | 0.354545 | 0.012785 | 0.015545 | 0.022034 | 0.080799 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.079934 | 0.000929 | 1 | 0.172727 | 0.186012 | 0.038419 | 0.159322 | 0.115297 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.262464 | 0 | 0 | 0.172727 | 0.025069 | 0.055741 | 0.022034 | 0.030413 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.268232 | 0.002323 | 0 | 0.218182 | 0.024066 | 0.007551 | 0.072881 | 0.001816 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.358467 | 0.000465 | 1 | 0.2 | 0.101529 | 0.030424 | 0.123729 | 0.065365 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.061805 | 0.003717 | 0 | 0.254545 | 0.059163 | 0.005108 | 0.020339 | 0.049478 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.247219 | 0.002788 | 0 | 0.181818 | 0.065681 | 0.000888 | 0.023729 | 0.154335 |

**Figure A2.** Sample of input data after one hot encoding and data transformation.

**Table A1.** Back propagation neural network training results.

| | Scheme 1 (No K-fold cross-validation) | | Scheme 2 (K-Fold cross-validation) | |
|---|---|---|---|---|
| Parameters | Test 1 | Test 2 | Test 1 | Test 2 |
| Number of training samples | (75:25) | (67:33) | 10Fold (90:10) | 5Fold (80:20) |
| Number of Hidden Neurons | 15 | 30 | 15 | 30 |
| Learning rate | 0.001 | 0.0001 | 0.001 | 0.0001 |
| Maximum Epochs | 1,000 | 1,000 | 1,000 | 1,000 |
| Training Time (s) | 33.79 | 21 | 188 | 82 |
| Loss | 0.3654 | 0.3675 | 0.3631 | 0.3650 |
| Training Method | Adam | Adam | Adam | Adam |
| Hidden Layer Activation Function | ReLU | ReLU | ReLU | ReLU |
| Hidden Layer Activation Function | Sigmoid | Sigmoid | Sigmoid | Sigmoid |
| Dropout percentage | 0.3 | 0.3 | 0.3 | 0.3 |

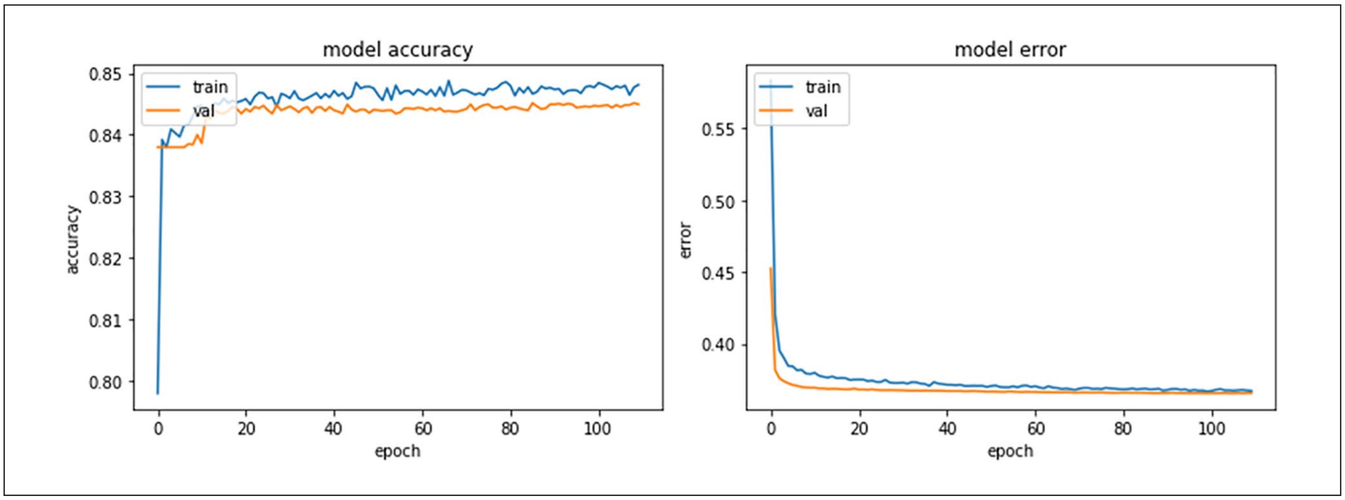


**Figure A3.** Accuracy and Loss Scheme 1 Test1 (75:25 Split).

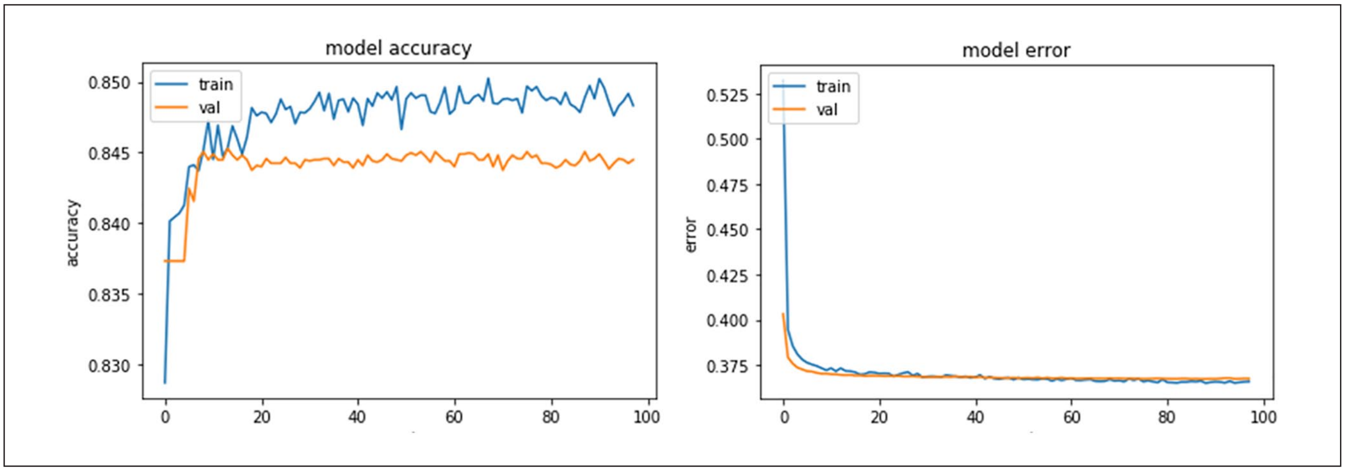


**Figure A4.** Accuracy and Loss Scheme 1 Test 2 (67:33 Split).

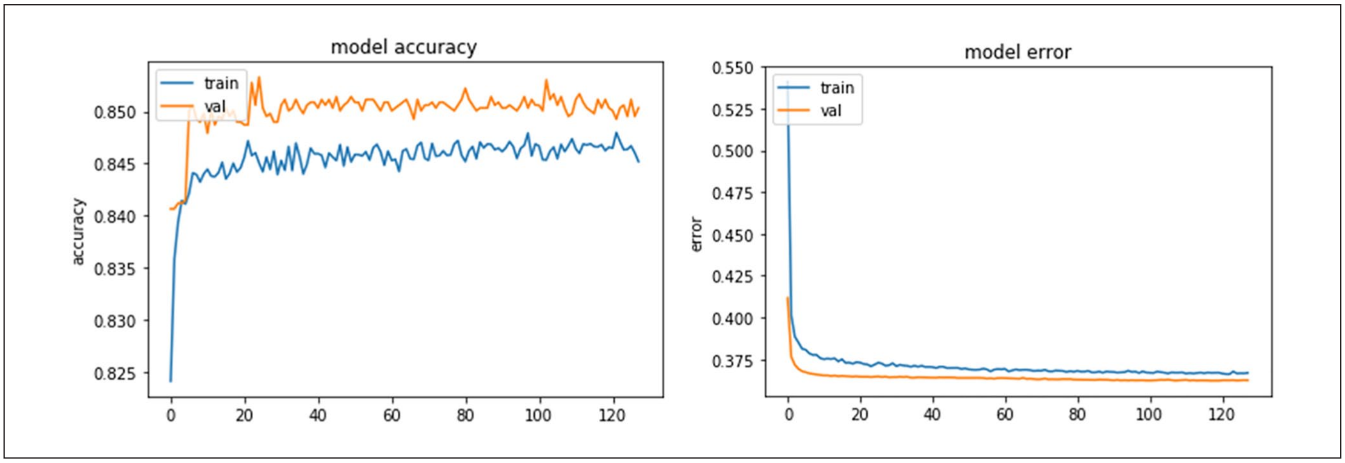


**Figure A5.** Accuracy and Loss Scheme 2 Test 1 (K-10 fold).

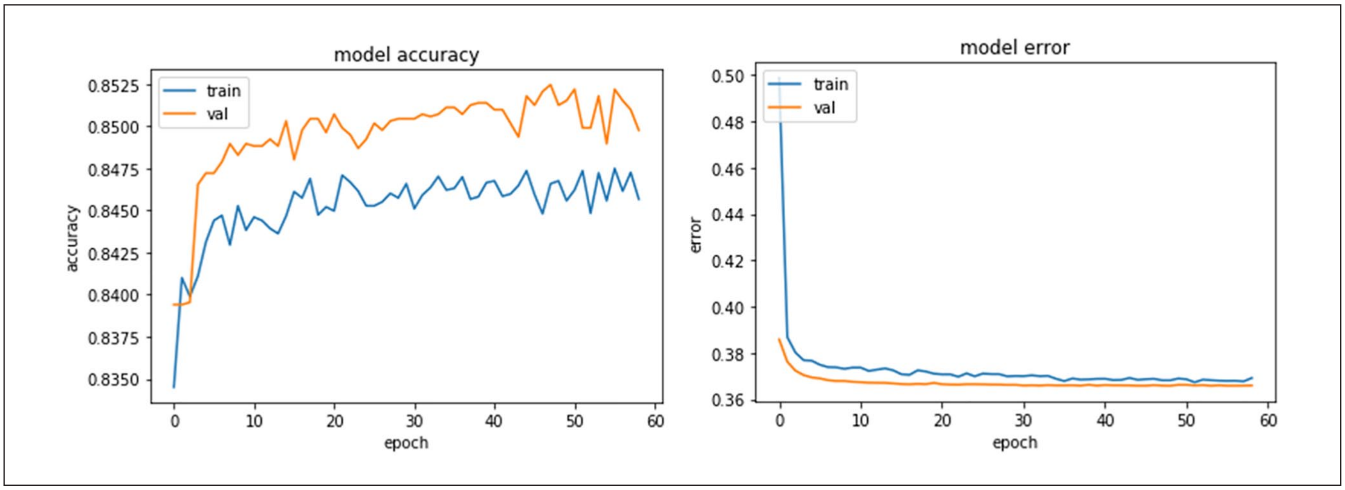


**Figure A6.** Accuracy and Loss Scheme 2 test 2 (K-5 fold).

**Table A2.** Back Propagation Neural Network Testing Results.

| | Scheme 1 (No K-fold cross-validation) | | Scheme 2 (K-10 fold cross-validation) | |
|---|---|---|---|---|
| Network parameter | Test 1 | Test 2 | Test 1 | Test 2 |
| Number of training samples | (75:25) | (67:33) | 10 fold (90:10) | 5 Fold (80:10) |
| Correctly Classified Training Samples | 4,720/5,578 | 4,221/4,983 | 5,678/6,695 | 5,046/5,951 |
| Recognition Rate on Training | 84.62% | 84.71% | 84.81% | 84.79% |
| Number of Test Samples | 1,860 | 2,455 | 743 | 1,487 |
| Correctly Classified Test Samples | 1,571 | 2,073 | 630 | 1,265 |
| Recognition Rate on Testing | 84.47% | 84.46% | 85.01% | 85.04% |
| Hamming Loss | 15.53 | 15.54 | 14.99 | 14.96 |
| Overall Recognition Rate | 84.54% | 84.58% | 84.84% | 84.92% |

## Acknowledgments

Authors would like to acknowledge David Stillwell, Michal Kosinski, and the myPersonality project in providing the myPersonality dataset for this study. First author is grateful to Mrs. Öcen Başaran for her contribution to the funding of this publication.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Ethical Declaration

The researchers declare that they did not collect any data from human/animal or any other subjects. Instead, the secondary data were used in this research that were retrieved from http://mypersonality.org/.

## Ethical Issues

This study uses secondary data; hence, all the participant information was anonymized by the providers of the dataset. Therefore, no consent was required.

## ORCID iD

Seren Başaran https://orcid.org/0000-0001-9983-1442

## References

Akshat, D. (2016). *Application of convolutional neural network models to personality prediction from social media images and citation prediction for academic papers* [Unpublished master's dissertation]. University of California, San Diego.

Al-Shihi, H., Sharma, S. K., & Sarrab, M. (2018). Neural network approach to predict mobile learning acceptance. *Education and Information Technologies*, *23*(5), 1805–1824.

Amichai-Hamburger, Y., & Vinitzky, G. (2010). Social network use and personality. *Computers in Human Behavior*, *26*(6), 1289–1295. https://doi.org/10.1016/j.chb.2010.03.018

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012). Personality and patterns of Facebook usage. In *Proceedings of the 3rd annual ACM Web science conference on" Websci'12* (pp. 24–32). Association for Computing Machinery Press. https://doi.org/10.1145/2380718.2380722

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, *21*(3), 372–374. https://doi.org/10.1177/0956797609360756

Bataineh, M., Marler, T., Abdel-Malek, K., & Arora, J. (2016). Neural network for dynamic human motion prediction. *Expert Systems with Applications*, *48*, 26–34.

Binh, H. T., & Duy, B. T. (2017). Predicting students' performance based on learning style by using artificial neural networks. In *Proceedings of the 2017 9th international conference on knowledge and systems engineering (KSE)* (pp. 48–53). Institute of Electrical and Electronics Engineers.

Champa, H., & AnandaKumar, K. (2010). Artificial neural network for human behavior prediction through handwriting analysis. *International Journal of Computer Applications*, *2*(2), 975–8887.

Chollet, F. (2017). *Deep learning with Python*. Manning Publications Co. https://dl.acm.org/citation.cfm?id=3203489

Corani, G., & Scanagatta, M. (2016). Air pollution prediction via multi-label classification. *Environmental Modelling & Software*, *80*, 259–264.

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, *13*(6), 653–665.

Devi, C., Reddy, B., & Kumar, K. (2012). ANN approach for weather prediction using back propagation. *International Journal of Engineering Trends and Technology*, *3*(1), 19–23. http://www.ijettjournal.org/volume-3/issue-1/IJETT-V3I1P204.pdf

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *Springer Plus*, *2*(1), Article 222. https://doi.org/10.1186/2193-1801-2-222

Espinosa, M. J., & Rodríguez, L. F. G. (2004). *Our Personalities: -In what and why we are dfferent*. Biblioteca Nueva.

Farnadi, G., Zoghbi, S., Moens, M., & De Cock, M. (2013). *Recognising personality traits using Facebook status updates* [Conference session]. Workshop on Computational Personality Recognition (WCPR13) in International AAAI Conference on Weblogs and Social Media (ICWSM13), Cambridge, MA, United States. https://biblio.ugent.be/publication/4237909

Golbeck, J., Robles, C., & Turner, K. (2011, May 7–12). *Predicting personality with social media* [Conference session]. Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems—CHI EA'11, Vancouver, British Columbia, Canada. https://www.cs.umd.edu/class/spring2017/cmsc818G/files/p253-golbeck.pdf

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Hall, J. W. (2017). *Examination of machine learning methods for multi-label classification of intellectual property documents* [Unpublished master's dissertation]. University of Illinois at Urbana-Champaign.

Heaton, J. (2008). *Introduction to neural networks with Java*. Heaton Research.

Hedberg, F., Granqvist, I., Nilsson, E., Skjutar, K., & Torstensson, P. (2010). *Predicting team performance based on artificial neural networks* [Defining the future of project management]. Project Management Institute. https://www.pmi.org/learning/library/team-performance-artificial-neural-networks-6494

Humphries, M. (2013). *Missing data & how to deal: An overview of missing data*. Population Research Center. https://liberalarts.utexas.edu/prc/_files/cs/Missing-Data.pdf

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, *2*, 102–138.

Kalghatgi, M. P., Ramannavar, M., & Dr Sidnal, N. S. (2015). A neural network approach to personality prediction based on the Big-Five model. *International Journal of Innovative Research in Advanced Engineering*, *2*(8), 56–63.

Kandias, M., Stavrou, V., Bozovic, N., & Gritzalis, D. (2013). Proactive insider threat detection through social media. In

*Proceedings of the 12th ACM workshop on workshop on privacy in the electronic society —WPES'13* (pp. 261–266). Association for Computing Machinery Press.

Kee, C. Y., Wong, L.-P., Khader, A. T., & Hassan, F. H. (2017). Multi-label classification of estimated time of arrival with ensemble neural networks in bus transportation network. In *Proceedings of the 2017 2nd IEEE international conference on intelligent transportation engineering (ICITE)* (pp. 150–154). Institute of Electrical and Electronics Engineers.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543–556. https://doi.org/10.1037/a0039210

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Laleh, A., & Shahram, R. (2017). Analyzing Facebook activities for personality recognition. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)* (pp. 960–964). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICMLA.2017.00-29

Lima, A. C. E. S., & de Castro, L. N. (2014). A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Networks*, *58*, 122–130. https://doi.org/10.1016/J.NEUNET

Liu, S. M., & Chen, J.-H. (2015). A multi-label classification based approach for sentiment classification. *Expert Systems with Applications*, *42*(3), 1083–1093.

Maxwell, A., Li, R., Yang, B., Weng, H., Ou, A., Hong, H., & Zhang, C. (2017). Deep learning architectures for multi-label classification of intelligent health risk prediction. *BMC Bioinformatics*, *18*(S14), Article 523. https://doi.org/10.1186/s12859-017-1898-z

McCrae, R. R., & John, O. P. (1992). An introduction to the Five-Factor Model and its applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

Mohammad, S. M., & Kiritchenko, S. (2013). Using nuances of emotion to identify personality. In *Proceedings of ICWSM* (pp. 1–4). https://arxiv.org/abs/1309.6352

Nam, J., Kim, J., Mencía, E. L., Gurevych, I., & Fürnkranz, J. (2014). Large-scale multi-label text classification: Revisiting neural networks. In T. Calders, F. Esposito, E. Hüllermeier, & R. Meo (Eds.), *Joint European conference on machine learning and knowledge discovery in databases* (pp. 437–452). Springer.

Nkoana, R. (2011). *Artificial neural network modelling of flood prediction and early warning* [Unpublished master's dissertation], University of the Free State.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://www.jmlr.org/papers/v12/pedregosa11a.html

Sharma, S. (2017). *Activation functions: Neural networks—Towards data science*. https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

Song, H. Y., & Kim, S. Y. (2014). Predicting human locations with Big Five personality and neural network. *Journal of Economics*, *2*(4), 273–280.

Statista. (2020). *Facebook users worldwide 2020*. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Sterbak, T. (2017). *Guide to multi-class multi-label classification with neural networks in python: Depends on the definition*. https://www.depends-on-the-definition.com/guide-to-multi-label-classification-with-neural-networks/

Sumner, C., Byers, A., & Shearing, M. (2011). Determining personality traits & privacy concerns from Facebook activity. *Black Hat Briefings*, *11*, 1–29. https://media.blackhat.com/bh-ad-11/Sumner/bh-ad-11-Sumner-Concerns_w_Facebook_WP.pdf

Tandera, T., Hendro Suhartono, D., Wongso, R., & Prasetio, Y. L. (2017). Personality prediction system from Facebook users. *Procedia Computer Science*, *116*, 604–611. https://doi.org/10.1016/j.procs.2017.10.016

Tareaf, R. B., Alhosseini, S. A., Berger, P., Hennig, P., & Meinel, C. (2019). Towards automatic personality prediction using Facebook likes metadata. In *Proceedings of the 2019 IEEE 14th international conference on intelligent systems and knowledge engineering (ISKE)* (pp. 715–719). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/iske47853.2019.9170375

Tsoumakas, G., & Ioannis, K. (2006). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, *3*(3), 1–13.

Wald, R., Khoshgoftaar, T. M., Napolitano, A., & Sumner, C. (2012). Using Twitter content to predict psychopathy. In *Proceedings of the 2012 11th international conference on machine learning and applications* (pp. 394–401). Institute of Electrical and Electronics Engineers.

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, *7*(3), 203–220.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), 2839–2846. https://doi.org/10.1016/J.PATCOG.2015.03.009

Xue, D., Wu, L., Hong, Z., Guo, S., Gao, L., Wu, Z., Zhong, X., & Sun, J. (2018). Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, *48*, 4232–4246. https://doi.org/10.1007/s10489-018-1212-4

Zhang, J. (2016). *Deep learning for multi-label scene classification* [Unpublished master's dissertation]. University of Adelaide.

Zhang, M.-L., & Zhou, Z.-H. (2006). Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, *18*(10), 1338–1351. https://doi.org/10.1109/TKDE.2006.162

Zhang, M.-L., & Zhou, Z.-H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048. https://doi.org/10.1016/J.PATCOG.2006.12.019

Zhu, Y. (2020). The prediction model of personality in social networks by using data mining deep learning algorithm and random walk model. *The International Journal of Electrical Engineering & Education*, 1–14. https://doi.org/10.1177/0020720920936839