Grenze International Journal of Engineering and Technology, June Issue

Mental Illness Prediction using Myers Briggs Personality Indicators

Venkatanaga Phalgunamanjush Venuthurumilli[1], Hrutuja Kargirwar[2], Kaustubh Shah[3] and Madhuri Sahu[4]
[1]G H Raisoni College of Engineering/Department, Nagpur, India
Email: manjushphalgun@gmail.com
[2-4]G H Raisoni College of Engineering/Artificial Intelligence, Nagpur, India
Email: {01hrutuja, kaustubhankushshah, madhuridec9}@gmail.com

***Abstract*—Personality has a significant weightage in our lives. It determines our cognitive abilities which directly affect our behaviour and our preferences and even affects our mental health. This project aims to predict users' mental health and MBTI personality using text inserted by the users via a web application. Utilizing Label Encoder and Count vectorizer, data was encoded. The machine learning model with the greatest accuracy score has been chosen. Several machine learning models were created and trained on the data-set, including Logistic Regression, SVM, Random Forest, Naive Bayes, and K-Neighbors Classifier. The XG-Boost classifier predicts 16 different personality types across four axes based on introversion, intuition, reasoning, and perceived abilities, even when tested on the same dataset. The accuracy which was predicted given by k fold Cross Validation shows that the XG-Boost classifier gives the best accuracy of 67.68 percent for prediction of the personality.**

***Index Terms*— Personality, MBTI, XG-Boost, SVM, Random Forest, Naive Bayes, K-Neighbours Classifier.**

I. INTRODUCTION

Our lives are constantly influenced by our personalities, from our behaviour to our choices in careers. By using personality features, it is feasible to create more precise recommendation systems and more effective digital marketing campaigns. Worldwide, rates of mental disease are soaring, and according to a WHO prediction, one in four people may have mental or neurological illnesses at some point in their lives. By 2020, depressive disorders will surpass all other diseases and overtake ischemic heart disease as the second greatest source of disease burden globally [1]. One of the most well-known and often-used personality type indicators is the Myers-Briggs Type Indicator (MBTI). With four binary categories and a total of 16 types, it depicts how people act and engage with the environment around them. Extrovert vs. introvert energy; sensing vs. intuition; thinking vs. feeling; and lifestyle vs. perceiving [2]. In this paper, we develop a text-based MBTI personality predictor that can identify mental illnesses. Based on the user's input text and a few answered questions, this model can first estimate the user's personality and then forecast their mental health.

II. RELATED WORK

The Myers-Briggs Type Indicator (MBTI) is a questionnaire-based personality profiling theory that aims to highlight psychological differences between individuals' perceptions of the world and decision-making skills.

Grenze ID: 01.GIJET.9.2.12

This scale was developed to distinguish between normal populations and naturally occurring individuals. The MBTI scale offers a wide variety of personality classifications by offering four personality classes. These four classes are further divided into sixteen personality types.

The four MBTI classes are: Extraversion vs Introversion: This class indicates how outgoing a person is. Extraverted people are talkative and outgoing. Introverted people are reserved, Sensing vs Intuition: This class indicates how a person perceives information. It differentiates between people who focus more on trusting what is certain and people who prefer to make inferences based on patterns and impressions, Thinking vs Feeling: This class indicates the decision-making ability of the person. An individual can either use logical reasoning or be more empathetic in making decisions and Judging vs Perceiving: This class indicates the orientation of an individual to the outer world.

An individual may either prefer being settled and organized, or spontaneous and flexible. A combination of the above four classes provides 16 MBTI personality types as shown in Table no. 1

In this project, a personality profile of a person is developed by the study of material that person has produced, such as an essay, tweet, or blog post. The primary focuses of the research are the type of data collected, text preparation techniques, and machine learning algorithms employed to estimate personality scores. The comparison of various feature vector combinations and machine learning models, as well as the deployment of solutions, have been discussed. The techniques described in this article allowed for accuracy levels up to 88. The Myers-Briggs Personality Indicator is a well-renowned technique of unmasking personality and was ameliorated to identify a person's personality strengths preferences and types [3].It was inferred that user data such as posts, statuses, essays and blogs were analyzed and users' personalities resulted in 88Gradient boosting which is a popular ML algorithm primarily utilized for classification and regression problems [5]. Brown Lee describes the analogies of distinct machine learning algorithms such as Naive Bayes, Gradient Boosting and SVM which were performed using a data set provided by Google named Bangle. All these machine learning algorithms performed reasonably well with acceptable results and accuracy, but gradient boosting gave him the best accuracy of 76.95A machine learning approach called gradient boosting is mostly employed for classification and regression issues [4]. According to Brown Lee, the Google Bangle dataset has been used to compare various machine learning techniques, including Naive Bayes, Gradient Boosting and SVM. Gradient boosting has the highest accuracy of 76.95 out of this plethora of machine learning techniques, which have all demonstrated promising results with respectable accuracy [5].

A student's intuition about success may show that students who study harder usually achieve higher grade point averages. Past research on this topic supports these findings that high-grade points in high school are co-related with personality traits of being conscientious.[8]. Furthermore, in the LASSO regression analysis reported in the Smart-Grade Point Average study at Dartmouth College, conscientiousness was chosen as the exclusive long-term predictor of greater grade point average. But there is hardly any research on the extent to which one's personality determines academic accomplishments in contrast to other external factors.

The Myers-Briggs Type Indicator is a personality-depicting proposition based on questions that aim to emphasize psychological distinctions between people's worldviews and decision-making abilities. This scale was created to differentiate between those that occur naturally and those that are part of a regular population. The four personality types offered by the MBTI scale allow for a wide range of personality classifications. Sixteen different personality types are created from these four classes. The following four MBTI types: Extraversion vs. introversion: This comparison categorizes people according to how outgoing they are. Talkative and extroverted individuals are extroverted. Reserving is a trait of introverts. Intuition vs. Sensing: This category describes how a person interprets data. It distinguishes between those who place a greater emphasis on believing what is certain and those who favour drawing conclusions based on inferences.

TABLE I. DIFFERENT PERSONALITY TYPES

Extraversion	Intuition	Feeling	Perceiving
Introversion	Sensing	Thinking	Judging

III. PROPOSED METHODOLOGY

Research shows that introverts are more prone to depression than extroverts. Because the MBTI classifies people as either introverted or extroverted, the BDI (Beck's Depression Inventory) can be applied to introverts to assess their mental health. This self-administered test clears any doubts the person may have about their mental health and allows them to consult a mental health professional if they have any concerns.

Fig. 1. Proposed approach

In the year 1917, Katherine Cook Briggs and Isabel Briggs Myers laid the foundation for the MBTI marker. The purpose of developing these nameplates during World War II was to assist women in their career choices.[9].

A. Visualization and Data Gathering

The data set utilized was obtained from Kaggle named" MBTI data-set". The number of columns is 2 and the numbers of rows are 8675. The data set does not contain null values. The drawback was that all the values were textual and had to be converted to arithmetic format to train the Machine Learning model One column indicates the personality type, i.e., ISTJ, ISTP, ISFJ, ISFP, INFJ, INFP, INTJ, INTP, ESTP, ESTJ, ESFP, ESFJ, ENFP, ENFJ, ENTP and ENTJ, while the other column contains the unstructured script written by that specific personality type. The latter contains the unorganized script that the specific personality typewrote. In the figure below, we can infer that the column contains various unnecessary details that we do not need, such as URLs. But, before we perform some pre-processing on our data set we analyzed the data to look for patterns so that we could understand the drawbacks we have in our data.
The dataset is clearly imbalanced between different classes. You'll observe that certain types of personalities contain much more data than others. The most common personality of Kaggle users is I-N-F-P (Introverted Intuitive feeling Perception). (The number of appearances is 1832 times). However, based on user comments, we have come to the following conclusions: For now, we can presume that users that commented more frequently on different social platforms are highly introverted, perceptive and more emotional.

B. Data cleaning and preprocessing

Data preprocessing is a method where the unstructured data is transformed into an interpretable format. Raw data can be incomplete, redundant, inconsistent and noisy. Data pre-processing includes different steps that help transform incomplete data into a complete and meaningful form[10].
Steps necessary for this process are converting posts to lowercase, Remove hyper links, and Lemmatization.
It was concluded that these posts contained many hyperlinks, and it was assumed that the URL links would not provide real information about the personality of the user, we applied regular expressions to clean up these data sets [11]. Later, each string was transformed into lowercase to prevent any differences in strings.
The next step, lemmatization was performed. Lemmatization is the method of combining distinct forms of a word so that they can be analyzed as a singular element. Lemmatization is comparable to stemming, but the context is brought to the words, so we use that in the model instead.[9].
We also removed posts with less than 15 words because the sentence length was not long enough to draw conclusions. It was performed with regular expressions.
In the next step, instead of selecting all 16 types of personalities as a unique feature, we hence trained 4 classifiers individually to classify their personalities based on MBTI type.
We created 4 new columns for each type indicator and assigned binary values. This is useful for calculations and comparing the number of introverted and extroverted posts from a given entry in the labelled Kaggledata set. This is performed in order to traverse the data set for all the individual Personality-Indices of MBTI. The figure below shows the same.

C. Feature Engineering using Tf-idf:

The term tf - idf is an amalgamation of two different terms: inverse document and term frequency. Term frequency is a number that shows how many times a particular word occurs in a set of documents, while inverse document frequency is a value that shows how important a particular word is in a set of documents.[12]. Feature engineering done using Tf–idf evaluates the relevance/importance of words to documents in a corpus of documents while training individual classifiers it is very helpful for natural language processing as we can score words in ML algorithms. We vectorize using Count-Vectorizer and tf-idf Vectorizer to keep words un-effected that appear between 10-70 percent of posts. The tf-idf approach is used mainly in information recall. The score for tf-idf is determined using the eq (1) [9].

$$Tf\ idf(t, d, D) = tf(t, d).idf(t, D). \quad (1)$$

Where’D’ is the set of documents, and ’t’ is the term frequency of the document ’d’. Whereas term frequency is calculated by the eq. (2) given below.

$$tf(t, d) = loge1 + freq(t, d))]\ [idf(t, D) = log(N/count(dED : tEd)). \quad (2)$$

D. Training and Hyperparameter tuning

We train and test the models using 6 different models namely Random Forest, XG boost, Stochastic Gradient Descent and Logistic Regression, SVM, KNN.
We split the features as:
X: tf idf portrayal of user posts
Y: Binary representation of personality type
Individual models are trained using X and Y characteristics. From the above models, we can see that the
XG Boost shows comparatively good performance on average. That’s why it was chosen to build the prediction model of different personalities. This is beneficial because the XG Boost model can also be used to evaluate and analyze the model tests execution during training.
Hyperparameters are parameters that have a large impact on the model’s architecture. Parameters such as polynomial functions, maximum depth, minimum number of samples, and learning rate are hyperparameters.[13].

E. BDI Questioner

The Beck Depression Inventory-II (BDI-II) is currently one of the most widely used methods to assess depression in both research and clinical practice. Depression is a common mood disorder that affects ‘functioning individuals in a variety of domains. Today, more than 350 million people worldwide suffer from depression, making depression a major contributor to the global burden of disease. Depression stands out not only for its high prevalence but also due to the probability of associated relapse and recurrence. A profound analysis of a large open-source database found that introverts may be more prone to depression than extroverts.[14].

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...

Fig. 2. Raw Data

	type	posts	IE	NS	TF	JP
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...	1	1	0	1
1	ENTP	'I'm finding the lack of me in these posts ver...	0	1	1	0
2	INTP	'Good one _____ https://www.youtube.com/wat...	1	1	1	0
3	INTJ	'Dear INTP, I enjoyed our conversation the o...	1	1	1	1
4	ENTJ	'You're fired.|||That's another silly misconce...	0	1	1	1

Fig. 3. Pre-Processed

IV. Results and Discussion

The below Table no. 2 shows the final accuracy of each classifier after hyper-parameter tuning.
Our model performed well but failed to predict individual extraversion in different cases. This is due to the reason that the data used to train the model had more posts from introverts than from extroverts. Had the data

```
Using CountVectorizer :
10 feature names can be seen below
[(0, 'ability'), (1, 'able'), (2, 'absolutely'), (3, 'across'), (4, 'act'), (5, 'actio
n'), (6, 'actually'), (7, 'add'), (8, 'advice'), (9, 'afraid')]

Using Tf-idf :
Now the dataset size is as below
(8675, 595)
```

Fig. 4. Count Vectorizer

been evenly distributed, the results would have been different. the average accuracy of all classifiers combined is 74.4 per cent which is pretty good.

TABLE II. NOTE HOW TABLE IS LEFT ALIGNED

Extraverted — Introverted	Accuracy: 77.37
Sensing — Intuition	Accuracy: 85.85
Thinking — Feeling	Accuracy: 70.28
Perceiving — Judging	Accuracy: 64.37

```
Introversion (I) /  Extroversion (E):    1999  /  6676
Intuition (N) / Sensing (S):             1197  /  7478
Thinking (T) / Feeling (F):              4694  /  3981
Judging (J) / Perceiving (P):            5241  /  3434
```

Fig. 5.

V. CONCLUSION

Many different technology and intelligent systems are being introduced in the mental health department to predict psychological problems at an earlier stage Machine learning can be a useful tool in helping understand psychiatric disorders. It may also aid in distinguishing and classifying mental health problems among patients for further treatment. A plethora of ML algorithms are available to develop intelligent systems, and comparing and selecting the best out of those algorithms is necessary. It is concluded from research that the XGBoost Classifier model predicts 16 different types of personalities from 4 axes with the highest accuracy possible. BDI questionnaire is presented to those individuals whose MBTI indicator predicts Introvert which will further help individuals assess their probable levels of depression. The current dataset used is small, In the coming future, it could be trained and tested using a larger dataset to increase the accuracy. we can’t completely rely on this technology as its accuracy is not so high but could be used for individual assessment. Mental health varies from individual to individual and there is a lot to learn about human cognitive functions.

REFERENCES

[1] Sumathi, M.R. and Poorna, B., 2016. Prediction of mental health problems among children using machine learning techniques. International Journal of Advanced Computer Science and Applications, 7(1).

[2] Cui, B. and Qi, C., 2017. Survey analysis of machine learning methods for natural language processing for MBTI Personality Type Prediction.

[3] Bharadwaj, S., Sridhar, S., Choudhary, R. and Srinath, R., 2018, September. Persona traits identification based on Myers-Briggs Type Indicator (MBTI)-a text classification approach. In 2018 international conference on advances in computing, communications and informatics (ICACCI) (pp. 1076-1082). IEEE.

[4] Mushtaq, Z., Ashraf, S. and Sabahat, N., 2020, November. Predicting MBTI personality type with K-means clustering and gradient boosting. In 2020 IEEE 23rd International Multitopic Conference (INMIC) (pp. 1-5). IEEE.

[5] Brownlee, J., 2016. A gentle introduction to the gradient boosting algorithm for machine learning. Machine Learning Mastery, 21.

[6] Soumik, M.M.J., Farhavi, S.S.M., Eva, F., Sinha, T. and Alam, M.S., 2019, December. Employing machine learning techniques on sentiment analysis of google play store bangla reviews. In 2019 22nd International Conference on Computer and Information Technology (ICCIT) (pp. 1- 5). IEEE.
[7] Bharadwaj, S., Sridhar, S., Choudhary, R. and Srinath, R., 2018, September. Persona traits identification based on Myers-Briggs Type Indicator (MBTI)-a text classification approach. In 2018 international conference on advances in computing, communications and informatics (ICACCI) (pp. 1076-1082). IEEE.
[8] Wang, R., Harari, G., Hao, P., Zhou, X. and Campbell, A.T., 2015, September. SmartGPA: how smartphones can assess and predict academic performance of college students. In Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing (pp. 295-306).
[9] Mushtaq, Z., Ashraf, S. and Sabahat, N., 2020, November. Predicting MBTI personality type with K-means clustering and gradient boosting. In 2020 IEEE 23rd International Multitopic Conference (INMIC) (pp. 1-5). IEEE.
[10] Agarwal, V., 2015. Research on data preprocessing and categorization technique for smartphone review analysis. International Journal of Computer Applications, 131(4), pp.30-36.
[11] Cox, R., 2007. Regular expression matching can be simple and fast (but is slow in java, perl, php, python, ruby,...). URL: http://swtch. com/rsc/regexp/regexp1. html.
[12] Scott, W., 2020. TF-IDF for Document Ranking from scratch in python on real world dataset. Medium.
[13] Jordan, J., 2017. Hyperparameter tuning for machine learning models. Retrieved from: Jeremy Jordan: https://www. jeremyjordan. me/hyperparameter-tuning.
[14] Hakulinen, C., Elovainio, M., Pulkki-Raback, L., Virtanen, M., ˚ Kivimaki, M. and Jokela, M., 2015. Personality and depressive symp- ¨ toms: Individual participant meta-analysis of 10 cohort studies. Depression and anxiety, 32(7), pp.461-470.