2022 2nd Asian Conference on Innovation in Technology (ASIANCON)
Pune, India. Aug 26-28, 2022

Myers Briggs Type Indicator (MBTI) - Personality Prediction using Deep Learning

Hrutik Naik
Department Of Information Technology
St. Francis Institute of Technology
Mumbai, India
hrutupn@gmail.com

Shrumi Dedhia
Department Of Information Technology
St. Francis Institute of Technology
Mumbai, India
shrumi.dedhia@gmail.com

Ashwini Dubbewar
Department Of Information Technology
St. Francis Institute of Technology
Mumbai, India
ashwinisd.mail@gmail.com

Meet Joshi
Department Of Information Technology
St. Francis Institute of Technology
Mumbai, India
mdjoshi00@gmail.com

Vandana Patil
Department Of Information Technology
St. Francis Institute of Technology
Mumbai, India
vandanapatil@sfit.ac.in

2022 2nd Asian Conference on Innovation in Technology (ASIANCON) | 978-1-6654-6851-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ASIANCON55314.2022.9909077

***Abstract*—Lately, there has been a massive spike up in the number of social network users. There is a massive evolution of social media platforms which have led to massive data generation. With this available data, there are large variations of methods to define personality of the users' based on their social behavior and patterns. With the aid of machine learning model and datasets the main aim of this paper is to predict the Myers–Briggs type Indicator (MBTI) personality type of the twitter user. The Myers-Briggs type Indicator is probably the maximum widely used personality check within the world. The predictor will help to predict 1 of 16 different personalities of the user based on their Twitter account. The text is preprocessed to get clean. After tokenizing of the data, machine learning model - LSTM (Long Short Term Memory Networks) has been built. The predictor endeavors to get of the personality type, traits and careers suitability.**

***Index Terms*—MBTI, Machine Learning, Personality Predictor, Deep Learning, LSTM, Long Short Term Memory Networks.**

I. Introduction

Artificial Intelligence is a wing of computer technology by which we can lay out intelligent machines which could act, decide, and make selections like human beings. The purpose of AI is to improve computer operations which might be related to human information, for example, interpretation, studying, and investigating. As technology including AI continues to increase, they may have a great influence on our standard of living. It's only reasonable that everybody these days desires to hook up with AI generation somehow, may additionally it be as an end-user or pursuing a profession in Artificial Intelligence. Machine learning is a subdivision of AI that utilizes data to resolve tasks. Those solvers are trained models of data that analyze primarily based on the information delivered to them. This data is procured from probability theory and linear algebra. Machine learning and deep learning models use our data to analyze and automatically resolve predictive tasks. Now-a-days, Deep Learning, a Machine Learning technique, is used to analyse the Twitter comments. In recent years, information growth has escalated in accordance with the emergence of social media, mainly within the form of textual data types. As a result of the inherent ambiguities of natural languages, growing a powerful personality prediction version based on the textual message that users share on social media may be a meticulously hard piece of work. In this paper, we orient toward predicting the user's personality based on the sort of comments he/she uses on Twitter. The Myers–Briggs type Indicator (MBTI) is an introspective self-report questionnaire indicating diverging predilections in how people understand the world and make choices. The recommended system stipulates the Myers–Briggs kind Indicator with a view to categorize the character types in sixteen patterns through 4 dichotomies, specifically, (1) Introversion - Extroversion, (2) Sensing - Intuition, (3) Thinking - Feeling and (4) Judging - Perceiving. A primary alphabet from each category or dimension will be taken into consideration to achieve a four letter test case, for instance, ISTP or ENTP or INFJ. The 4 letter case helps to acquire the correct outcomes. Each kind is stated to outline a specific set of behavioral dispositions, reflecting variations in attitudes, orientation, and choice-making patterns. As an example, the individual could be appropriate for the job position of Analysts, if the output class of 4 letter case might be acquired as INTJ - a personality type which describes people as innovative and strategic thinkers, with a plan for everything. The MBTI personality type is predicted using Deep Learning model (DL) - LSTM (Long Short Term Memory Networks). It is difficult to categorize the personality and job suitability

of the person, so the intention of the system is to annihilate this barrier and give out the personality types throughout 4 dichotomies which suggest the psychological preferences and aids to provide job suitability of the person. [1]

II. Literature Review

A. Literature review related to existing methodology

In [1], personality was predicted by using NLP, CNN, LSTM, LSG algorithms. Firstly each word was embedded into the word vector and encoded in the first part. In the second part, 2 LSTM (Long Short Term Networks) and CNN (Convolutional Neural Network) are used to learn the structural features of the output from the first part. In the final part concept of LSG (Latent Science Group) is used, the output from the second part is sent into softmax to produce the personality traits as the final result.

HRV gives critical mental health records for medical diagnosis, telemedicine, preventative medication, and public health. but, the dearth of a realistic detection mechanism restricts its application. The purpose of this is to have a look at [2] changed into to observe the feasibility and reliability of using smartphone Photoplethysmogram (PPG)-based totally HRV analysis for personality prediction. In Shenzhen, China, 95 comments were amassed from college students and college personnel. An app took 5-minute films of their arms and turned the frames into HRV measurements. individuals who had been more extraverted and stable had a extra root imply square of successive variations (rMSSD; p=0.03 and 0.005, respectively), as well as a bigger percent of consecutive ordinary-to-normal (NN) durations that numerous by extra than 50 ms (pNN50; p=0.05 and 0.004, respectively), and SDNN (p=0.02 and 0.01, respectively). stable humans also had better log high-frequency HRV (p=0.008). The correlation coefficients and the bland–Altman evaluation outcomes verified the accuracy of cellphone PPG in HRV measurement. All HRV values acquired utilizing smartphone PPG and reference ECG had correlation values greater than 0.9. Moreover, for all HRV measurements besides pNN50, the bland–Altman ratios had been less than 0.2. Taken collectively, the findings of this have given the first empirical records that supported the use of cellphone PPG as a personality predictor.

The quantity of human beings and the usage of social media has skyrocketed in recent years. In this context [3], social media furnished researchers with a wealth of facts about user and societal conduct. They were starting to realize how a person's conduct on social media pertains to their personalities. Conventional personality assessments rely upon self-document inventories that are expensive to accumulate. These studies tried to predict a user's big-five character based on information obtained from social networks. They administered a large-five personality stock exam to 131 Sina Weibo users and extracted all of their Weibo messages and profile facts. The authors correctly expected the big-5 personality of customers by means of investigating the relevance between all forms of consumer produced facts and personal results of customers. They extracted traits which include consumer behavior, interplay conduct and textual content language conduct. Authors used Pearson Correlation Coefficient to select functions based totally on dependency metrics concept. They used machine learning knowledge of algorithms to expect rankings of personalities with extracted capabilities.On this paper, Logistic Regression and Naïve Bayes algorithms were used to get the large-5 personalities. They decided on 5 maximum relative dimensions and applied a machine learning algorithm to know the approach to correctly forecast the big-five personalities of users.

They [4] acquire social data and questionnaires from Weibo users and concentrate on how to leverage user text information to predict personality traits. The authors used correlation analysis and principal component analysis to choose the user data, and then used the multiple regression model, grey prediction model, and multitasking model to predict and evaluate the outcomes. The grey prediction's MAE values were found to be better than the Multitask model's multiple regression model, with the total effect of the prediction ranging between 0.8 and 0.9, showing good prediction accuracy. The grey prediction model exhibited superior prediction accuracy than the other two regression models, according to the MAE prediction index.

In [5],user behavior at the facebook social networking web page became used to make personality predictions. With the emergence of social networks, a slew of latest strategies for figuring out people's personalities based totally on their social activities and linguistic styles have arisen. Machine Learning algorithms, records sources, and feature sets fluctuate between methodologies. The intention of this examination turned into to peer how well massive five version characteristics and metrics might expect facebook customers' persona traits. Tokenization became employed to do away with URLs, symbols, names, and lowercase letters. They inferred a person's traits from functions related to a person's social network.. The Pearson correlation analysis turned into hired as the standard function selection method to quantify the power of the linear dating among two variables and to take a look at capabilities widespread for personality trait prediction. The effects of the prediction accuracy tests tested that the personality prediction system based at the XGBoost classifier outperformed the common baseline for all characteristic units, with the very best prediction accuracy of seventy 74.2 Percentage. With a prediction accuracy of 78.6 Percentage, the character social community evaluation functions set accomplished the exceptional prediction overall performance for the extraversion feature. information pre-processing, feature extraction, and function choice are all steps inside the method.

B. Literature Review related to existing Algorithms

In [1], Algorithms such as NLP (Natural Language Processing), CNN (Convolutional Neural Network), LSTM (Long Short Term Memory), LSG (Latent Science Group) were employed. Authors in [2] used Spearman correlation coefficient algorithm and Bland–Altman method. Logistic Regression and Naïve Bayes algorithms were utilized in [3]. Here [4] researchers used the Multiple Regression Model, Multitask Regression Model and Grey Prediction Model. In this paper [5], XGboost algorithm as the primary classier Support Vector Machine (SVM), Logistic Regression and Gradient Boosting as a baseline for comparison were used by authors.

C. Literature Review related to Tools/Technology/framework

In [1], The authors used Weka software. Weka is a library of machine learning algorithms for use in data mining jobs. In [2], tools such as Electrocardiogram (ECG), Camera-based PPG, and HRD were used. An electrocardiogram (ECG or EKG) is a test that measures the electrical signal from your heart in order to diagnose various cardiac diseases. Electrodes are implanted on your chest to capture the electrical signals produced by your heart, which cause it to beat. On an associated computer monitor or printer, the signals are displayed as waves. PPG (photoplethysmography) sensors use a light-based technology to sense the rate of blood flow as controlled by the heart's pumping action. The authors had used Weka in [3]. In this paper [5], NLTK was used by authors. The Natural Language Toolkit (NLTK) is a Python framework for developing programmes that interact with human language data and may be utilized in statistical natural language processing (NLP). Text processing packages for tokenization, parsing, categorization, stemming, tagging, and semantic reasoning are included.

D. Gaps Identified

- The occurrence of the randomness occurring due to individual differences and various environmental variables will be removed with the use of a within-group design thus increasing the statistical power of the comparison.
- Recruiting more participants and collecting more information will give a larger dataset. So, the precision of prediction of five personalities can be improved.
- Instead of collecting data from a specific social network, other more widely used social media will help to get better dataset and more information as well as questionnaires will be used for precise prediction of personality.

III. Proposed Methodology

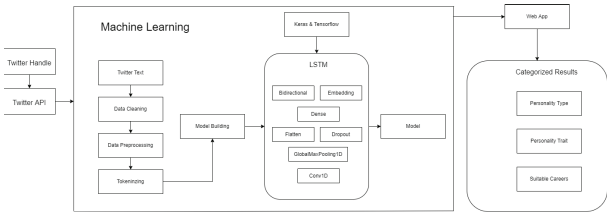

Fig. 1. Architectural Design

A. Problem Formulation

The proposed system aims to address how the dissatisfying accuracy of personality type prediction based on MBTI can be improved as well as it addresses the need to have larger dataset so that classification can be performed for all dichotomies.

B. Problem Definition

Many people find it difficult to identify the suitable career for themselves. This leads to them making wrong decisions and choosing a wrong career path

C. Scope

The aim of the Myers-Briggs Type Indicator (MBTI) personality predictor is to make the social media that is Twitter texts understandable to the users in identifying their personality type which is useful in determining their job suitability. The purpose is to help the user in recognizing any trends or patterns that can be detected in their style of writing Twitter texts, which in turn helps in analyzing, predicting or categorizing behavior into different dichotomies. The predictor helps the users to identify if they are Analyst or Diplomats or Sentinels or Explores. By analyzing candidates' Twitter texts, MBTI personality types gets predicted.

D. Proposed Methodology

At the beginning, the proposed system asks the users to enter their twitter handle. Then the Twitter API is used to scrap the data of the users. Firstly the system does the preprocessing of the Twitter texts. In this phase of preprocessing, the cleaning of the text is done. It includes removing all special characters and numbers. All the letters transforms into lowercase. All the sentences is tokenized. Tokenization will help in returning the more complete root words. After data gathering and data pre processing, the machine learning model - LSTM is created. Then the data is categorized in four dichotomies when classifying MBTI types, which will give 16 distinct outputs. Finally, the user obtains the categorized results in form of their MBTI personality type with mapping chart, personality traits and their career options.

E. Proposed Model

Long Short-Term Memory Networks (LSTM): Apparently, Deep Learning Models have been in focus to get personality prediction. Subjectively, the convolutional neural network (CNN) strives to recreate the process of creating articles, whereas the recurrent neural network (RNN) attempts to analyse texts by mimicking the human reading process.

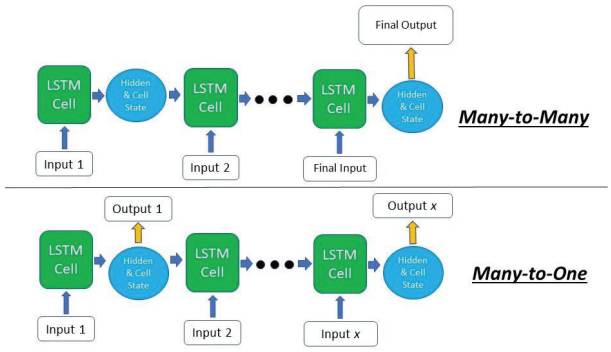

Fig. 2. RNN: Architectural Design

In RNN [13], initially, words are converted into machine-readable vectors. The RNN then goes over the vector sequence one by one. It sends the previous hidden state to the next stage of the sequence while processing. The neural network's memory is stored in the hidden state. It stores information about prior data that the network has seen. First, the prior hidden state and the input are merged to generate a vector. That vector now contains knowledge regarding the current and prior inputs. The vector is activated by tanh (an activation function), and the result is the network's new hidden state or memory.

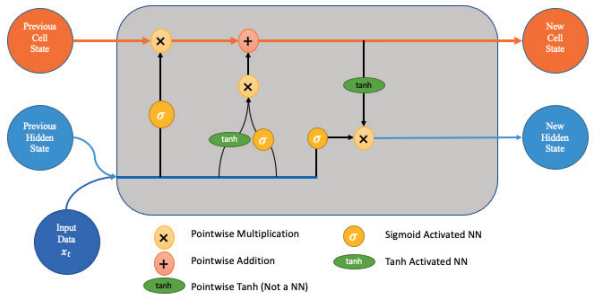

Fig. 3. LSTM: Architecture

The Long Short-Term Memory Networks (LSTM) model [13] is used in this project. The concept of an LSTM is similar to that of a recurrent neural network. It processes data and forwards data as it propagates. The processes within the LSTM's cells are what distinguishes them. These processes allow the LSTM to remember or forget information. LSTM is made up of a cell, that is, forget gate, input gate and ouput gate. The core concept and cells' operations of LSTM are explained below:

- Tanh: Tanh is an activation function that is not linear. It manages the values that pass over the network, keeping them between -1 and 1. To avoid information fading, a function with a longer second derivative is required. It is possible that certain values will become gigantic, leading other values to become trivial. Because of the function, the number 5 persists between the limits.
- Sigmoid: The sigmoid function is a type of non-linear activation function. The gate keeps it at bay. Sigmoid, unlike tanh, keeps values between 0 and 1. It assists the network in updating or forgetting data. If the multiplication yields a 0 result, the information is deemed lost. Similarly, if the value is 1, the information remains.This will assist the network in determining which data may be forgotten and which data must be retained.
- Forget Gate: The forget gate determines which information is important and which may be disregarded. The sigmoid function is used to process data from the current input and the hidden state. The Sigmoid function creates values between 0 and 1. It determines if a portion of the former output is required (by giving the output closer to 1). The cell will eventually utilise this value for point-by-point multiplication.
- Input Gate: To update the cell state, the input gate conducts processes. First, the second sigmoid function receives the current state and the previously hidden state. The values are changed between 0 (important) and 1 (unimportant). The identical information about the hidden and current states will then be supplied through the tanh function. To control the network, the tanh operator will generate a vector containing all possible values between -1 and 1. The activation functions' output values are ready for point-by-point multiplication.
- Output Gate: The value of the next hidden state is determined by the output gate. This state holds data from earlier inputs. First, the current and prior hidden state values are supplied into the third sigmoid function. The new cell state that is created from the cell state is then sent to the tanh function. These two outputs are multiplied point by point. The network determines which information the hidden state should carry based on the final value. This concealed state is employed in prediction. Finally, the new cell state and concealed state are passed forward to the following time step.

F. Features of Proposed System

- The system Predicts distinct personality types.
- The system helps the users explore how their personality traits affect various aspects of their lives.
- The system offers career suitability of the user.

IV. Implementation

A. Working of the Project

The proposed system uses machine learning to enhance the user experience for getting their MBTI personality type

with suitable job or employment status. The project uses LSTM model which helps in creating the personality type that is identical to the user's personality as it is in real life. Implementation involves the following steps:

Step 1: Data Gathering: First the data is gathered for which the project is to be made. According to the data gathered, the models are designed. The Twitter texts are gathered and the preprocessed.

Step 2: Data Preprocessing: In this stage, data is preprocessed in the text filtering stage so that the machine learning model in the next stage can interpret it. The input data for Twitter messages includes Uniform Resource Locators (URLs), numbers, foreign language words, abbreviations, symbols, and emoticons. To eliminate all of these unnecessary characters from the supplied data, data preparation is performed. Following data preparation, the input data is classified based on the varied demands of the individuals.

Step 3: Model Creation: As this project deals with the machine learning models, a model is created that helps to categorized and determine the MBTI personality type of the users.

Step 4: Model analysis: The Deep learning model - LSTM is developed in order to get the accurate results of the personality prediction.

Step 5: Comments analysis: Once the model has been developed, the comments are fed to analyze the Personality type predictor which uses LSTM to predict the personality.

Step 6: Coding the Functions: All the remaining necessary Machine Learning functions which are required to predict the user's personality are implemented using Python.

Step 7: Building the Web Application: Once the machine learning code component is done, a web application is created to analyze the input comment and determine the proper personality type with job appropriateness. Flask is used to connect the back end with the web application created.

Step 8: Testing: The various comments are entered here to see if the input data produces the proper output, i.e. the correct MBTI personality of the user is predicted on the basis of the Tweets.

Step 9: Deployment: Finally, the project is made available for users to test and determine their personalities. The predictor helps to predict 1 of 16 different personalities of the user along with traits and job suitability based on their Twitter account.

B. Application Performance

The user requires a laptop with the website that works as follows:

- Input Data: When the user enters the comment to be evaluated for personality prediction and submits it by clicking on the button on the home page the comment gets processed and tokenized.
- Trait Classification: After processing the comments, out of 16 dichotomies, 4 dichotomies are classified together and sent to the predictor.
- Personality Prediction: The predictor provides us with a personality type which contains 4 traits of personality out of the 16 MBTI personality traits. Using this it can be predicted if the candidate is eligible for the job or not.
- User Interaction: The user is provided with several on-click functionalities to interact with the page and view additional information. All the 16 traits are well explained once the personality is predicted. A brief explanation about every job role is provided by highlighting the key qualities of that person.

C. Results

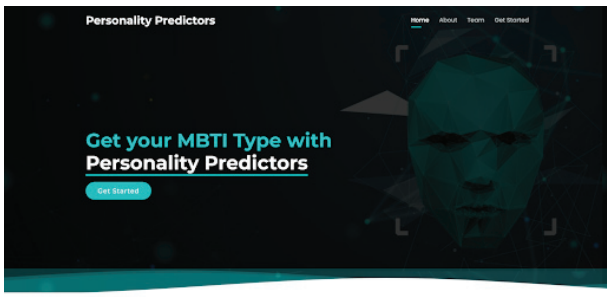

Fig. 4. Home Screen.

Fig. 5. Twitter Handle: Input to get MBTI Personality Type

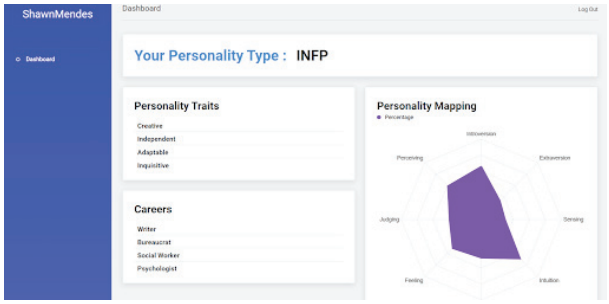

Fig. 6. Output 1: Personality Type, Personality Traits, Career Suitability along with mapping function

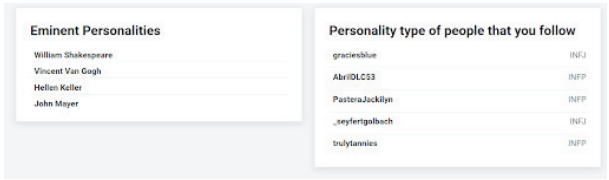

Fig. 7. Output 2: Similar Personality and Personality of the followers

V. Conclusion

The study investigates the literature on the uses of social media framework as behavioral features by exploring the relationship between users' personalities and their behaviors in social networks. To predict a user's personality, we have conducted a study of best behavioral indicators. To conduct our research, we have utilized a dataset to assess several personality factors that indicates job fit.

Acknowledgement

We would like to acknowledge our indebtedness and render our warmest thanks to our mentor, Ms.Vandana Patil, who made this work possible. Her friendly guidance have been invaluable throughout all stages of the work. This thesis has been written during the Bachelors course in Information Technology at St.Francis Institute of Technology. We would like to thank SFIT for providing excellent working conditions and for its support.

References

[1] X. Sun, B. Liu, J. Cao, J. Luo and X. Shen, "Who Am I? Personality Detection Based on Deep Learning for Texts," 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 2018.

[2] I. Liu, S. Ni and K. Peng, "Predicting Personality with Smartphone Cameras: A Pilot Study," 2020 IEEE International Conference on Human-Machine Systems (ICHMS), Rome, Italy, 2020.

[3] Wan D., Zhang C., Wu M., An Z. (2014) "Personality Prediction Based on All Characters of User Social Media Information". In: Huang H., Liu T., Zhang HP., Tang J. (eds) Social Media Processing. SMP 2014. Communications in Computer and Information Science, vol 489. Springer, Berlin, Heidelberg, 2014.

[4] C. Li, J. Wan and B. Wang, "Personality Prediction of Social Network Users," 2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES), Anyang, China, 2017.

[5] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Personality Predictions Based on User Behavior on the Facebook Social Media Platform," in IEEE Access, vol. 6, pp. 61959-61969, 2018.

[6] T. Tandera, D. Suhartono, R. Wongso, Y. L. Prasetio et al., "Personality prediction system from facebook users," Procedia Computer Science, vol. 116, pp. 604–611, 2017.

[7] V. Ong, A. D. Rahmanto, D. Suhartono, A. E. Nugroho, E. W. Andangsari, M. N. Suprayogi et al., "Personality prediction based on twitter information in bahasa indonesia," in Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on. IEEE, 2017.

[8] V. Evrim and A. Awwal, "Effect of personality traits on classification of political orientation," World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, vol. 9, no. 6, pp. 2001–2006, 2015.

[9] H. Wei, F. Zhang, N. J. Yuan, C. Cao, H. Fu, X. Xie, Y. Rui, and W.-Y. Ma, "Beyond the words: Predicting user personality from heterogeneous information," in Proceedings of the tenth ACM international conference on web search and data mining. ACM, 2017, pp. 305–314.

[10] J. Maria Balmaceda, S. Schiaffino, and D. Godoy, "How do personality traits affect communication among users in online social networks?" Online Information Review, vol. 38, no. 1, pp. 136–153, 2014.

[11] J. Golbeck, C. Robles, M. Edmondson and K. Turner, "Predicting Personality from Twitter," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011, pp. 149-156, doi: 10.1109/PASSAT/SocialCom.2011.33.

[12] Li, L., Li, A., Hao, B., Guan, Z., Zhu, T.: Predicting Active Users' Personality Based on Micro-Blogging Behaviors. PLoS ONE 9(1), e84997 (2014), doi:10.1371/journal.pone.0084997

[13] https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21