International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

MBTI Based Personality Prediction of a User Based on Their Writing on Social Media

Neha Gupta, Anirudh Madhavan, Divya Duvvuri, R.Angeline

Abstract*: The undeniable power that various e-commerce and streaming websites exert on their users' in terms of what they buy and what they watch is unquestionable, so the creation of better targeted advertisements and recommender systems is the need of the hour. Prediction of a person's personality can be the key for the achievement of these goals. A novel way to understand the various facets of a person's personality is by analyzing their MBTI (Myers–Briggs Type Indicator). This paper aims at classifying a user into any one of the sixteen personality types, defined by MBTI, through the use of natural language processing (NLP) and support vector machine (SVM) which was implemented on the MBTI dataset. Since the original dataset is unevenly distributed, SVM has been applied to the original dataset and an under sampled version of the MBTI dataset. The highest accuracy rate of 78.52% for the traits (thinking/feeling) was achieved in the original dataset whereas for the under sampled dataset it was 60.2% for the traits (judging/perceiving).

***Keywords*: MBTI, NLP, Personality Prediction, SVM**

I. INTRODUCTION

The personality of a person defines the way they respond, understand and process information. Numerous personality tests such as the Big five model, Eysenck Personality Inventory, DISC etc., are used to break down the complexities of a person's personality into various factors which can then be easily scrutinized. A user's activity on social media, particularly in the way they express themselves based on their writing is a powerful method to understand the various nuances of a person's personality. Each person has their own unique outlook on life which when tapped into, can create a more personalized environment for the user. Carl Jung developed the "Theory of psychological type" which proposes the existence of two pairs of traits, one indicating how a person's energy is either internalized or externalized (introversion or extraversion) and the other one defining the way they respond to information (judging or perceiving). Isabel Briggs Meyers and her mother Katherine Cook Meyers then set out to implement Jung's ideas in a much more practical way, which lead to the creation of the Myers Briggs Type Indicator. According to the MBTI each person is identified by a unique four coded formula which is based on the analysis of a total of eight traits.

1. **E**xtraversion (E) or **I**ntroversion (I)
2. I**n**tuition (N) or **S**ensing (S)
3. **F**eeling (F) or **T**hinking (T)
4. **P**erceiving (P) or **J**udging (J)

Based on which trait is satisfied among the pair of traits, a four coded formula would be generated. For example: the code INFP indicates that the person has the following traits: introversion, intuition, feeling and perceiving. The traits along with their explanations are given in Table I.

Table- I: MBTI Traits and Their Explanations

Trait	**Explanation**
Introversion or Extraversion	Whether their energy is internalized or externalized
Sensing Intuition	How they process information
Thinking Feeling	Decision making skills
Judging Perceiving	Their response to the external world

Our study uses NLP and SVM to scrutinize the MBTI dataset [11], which consists of nearly 8,600 rows of unique data. In order to predict the four coded symbol, the SVM model is applied to each of the four trait pairs. On analyzing the dataset it was noted that data for the 'Extrovert' and 'Sensing' traits was highly imbalanced in comparison to their counterparts. In order to combat this imbalance, an under sampled version of the dataset was also created and the SVM model was applied to the newly created dataset.

The organization for the rest of the paper is as follows. In section II we discuss the work that was carried out related to personality prediction. In section III we discuss about the dataset and its visualization. In section IV we describe the methodology that was applied. In section V the result is discussed, followed by section VI where we conclude our study and explain about future work.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Neha Gupta*, Computer Science and Engineering, SRM institute of science and technology, Chennai, India.
Email: neha_gupta17@srmuniv.edu.in
Anirudh Madhavan, Computer Science and Engineering, SRM institute of science and technology, Chennai, India.
Email: anirudh_ragavender@srmuniv.edu.in
Divya Duvvuri, Computer Science and Engineering, SRM institute of science and technology, Chennai, India.
Email: duvvuri_divya@srmuniv.edu.in

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

II. RELATED WORK

Research in personality prediction has increased exponentially in recent years. A major reason for this boost would be the numerous benefits of the determination of a user's personality. The vast majority of this work has been done using the Big 5 factor model. It categorizes people based on five traits, namely, openness, conscientiousness, extraversion, agreeableness and neuroticism. Tadesse *et al.* [1] classified a person's personality according to the big five model. Their findings show that users with extraversion trait write shorter and brief sentences and use words indicating positivity and emotion compared to neurotic users. Neurotic users, on the other hand, use words indicating negative emotion and write longer posts and sentences. Users having high conscientiousness, on the other hand, prefer talking about happy subjects. People with high openness use high frequency function words in contrast to content words. Agreeable users tend to use more question marks and interrogative sentences. Tasfia Hoque *et al.*[2] conducted a very unique survey, where they collected the reviews of a restaurant based on their food and then predicted the personality of a person based on those reviews. A questionnaire containing 22 questions was prepared which was plugged into the model that was being developed generating a 70% accuracy rate for 'judge/perceive' trait. Naveen K. Kambham *et al.* [3] predicted the personality type of the user directly through their mobile phones without any user interaction. Daily routines of participants were measured from readily available parameters from smartphones. Their paper is based after a well respected model of personality called the big five model. Backpropagation algorithm, which utilized root mean squared error (RMSE) as the metric of quality, was chosen as the training algorithm for the models. Guozhen Zhao *et al.* [4] modelled a paper which aims at recognizing personality traits by analyzing brain waves. Model classifies according to the big five frame work. Results of the model show that accuracy rates which are above 80% are generated for three personality traits namely agreeableness, extraversion and conscientiousness when emotions which are positive are evoked. On the other hand, classification accuracy for the neuroticism trait is 78%-81% when emotions which are negative, except for disgust, are evoked than positive emotions, and the accuracy of classification of the openness trait is 83% when a film clip which is disgusting is shown. Anisha Yata *et al.* [5] described a paper wherein the personality of a person was predicted through a personality test. Their paper talks about how the task of prediction of personality can be automated by the application of multi label classifiers on text based data. It uses various classification methods in addition to multi label classifiers which aided in the task of personality prediction. Tao Ding *et al.* [6] modeled a paper that classifies people accurately which would, in turn, enable the prediction of their personal preferences, ~~it~~ and would also help in the creation of better targeted advertisements. This was done through data that was compiled from the visitors of an amusement park. Through this paper, the choices that an individual makes were analyzed which shows the correlation of choices and characteristics. Results show that when models are trained with the features which are selected due to casual identification performed better than any existing structures. Junjie Chen *et al.* [7] modelled a paper that studies the allocation of a particular bandwidth depending upon the personality traits of a smartphone user. This would have the benefit of a much smarter use of mobile bandwidth. Paper utilizes characteristic inference, the probability of a particular personality trait existing in a user can be calculated by the service provider, depending upon the amount of data used. Mihal Gavrilescu *et al.* [8] proposed a novel way for the detection of personality traits based on the big five model. This could be done using neural networks which worked on handwriting features. Their paper shows that texts that are predefined can add more merit if applied to authors in the training stage, leading to accuracy rates of around 84% in the case of intra subject tests and 80% in inter subject tests, when testing is carried out using the random dataset. The accuracy rates are the highest for Experience, Openness, Neuroticism and Extraversion (over 84%), whereas for Agreeableness and Conscientiousness, the accuracy is 77%. Giulio Carducci *et al.* [9] modelled a paper that determines the various personality traits of an individual based on their tweets, this was done through a supervised learning method. The methodology that they followed first breaks the tweets down to numerous tokens through which it then learns the representation of various words as embeddings. This was then fed to a supervised classifier. Raad Bin Tareaf *et al.* [10] modelled a paper that utilized research that was done in the field of analysis of text and detection of personality to develop an 'automatic brand personality prediction model'. This was done based on the big five model, and features such as linguistic inquiry and word count were used from sources available publicly. The model that was proposed in the paper detailed high accuracy rates for the prediction of certain personality traits from brands.

III. DATA AND ITS VISUALISATION

To classify users' according to their MBTI we employed the MBTI dataset [11], which was collected from the my Personality forum. It contains 8,600 rows of unique data, with each row containing information specifying the type and posts, where type represents the four coded MBTI type followed by the first 50 posts that were made by the user of the specified type. Each post is separated by '| | |' (3 pipe characters). On analyzing the number of users' available for each trait it was observed that there was a significant imbalance between introverts and extroverts (6678 (I) VS 1999 (E)) in terms of number of users,and a similar situation was observed for the sensing and intuition traits (7478 (N) VS 1197 (S)). In contrast, the pairs feeling/thinking (4694 (F) VS 3981(T)) and judging/perceiving (5241 (J) VS 3434 (P)) were balanced. This is graphically represented in fig. 1 for I/E, fig. 2 for N/S, fig. 3 for F/T, fig. 4 for J/P. A highly unbalanced dataset would lead to inaccurate predictions as the majority class would be favored over the minority class. There are two ways to combat this problem: we could either under sample or oversample the dataset. Under sampling, as the name suggests involves scaling down the majority class to the minority class, while on the other hand, oversampling simply over samples the minority till it matches with the majority class. Since oversampling can lead to problems of overfitting we decided to try out the former. An under sampled version of the original dataset was created.

A pie chart depicting the parts of speech (POS) was created which aided in data visualization and enabled us to gain valuable insights regarding the similarities between every corresponding trait.

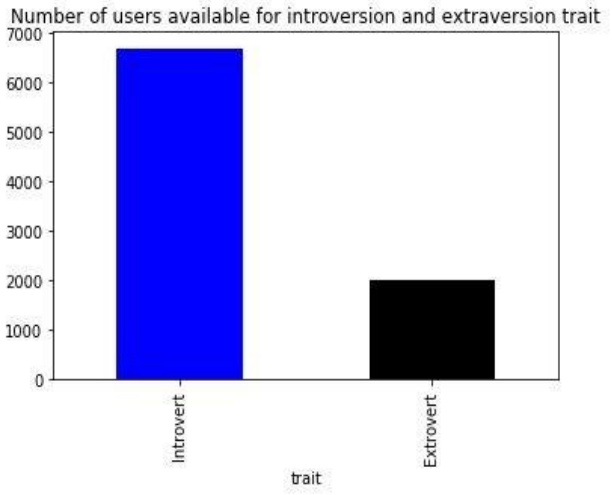

Fig. 1 User Count for Introversion VS Extroversion Trait

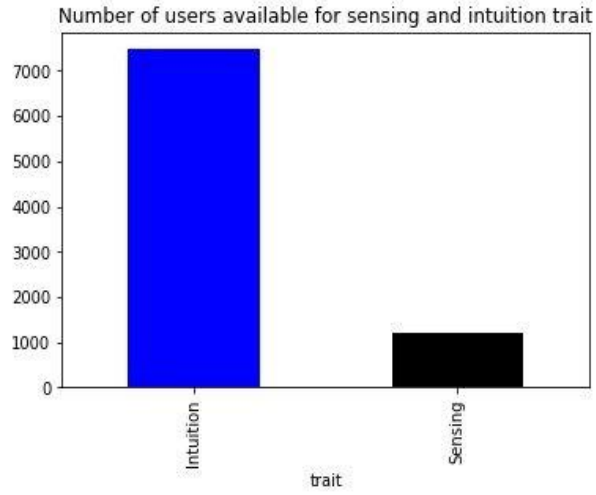

Fig. 2 User Count for Sensing VS Intuition Trait

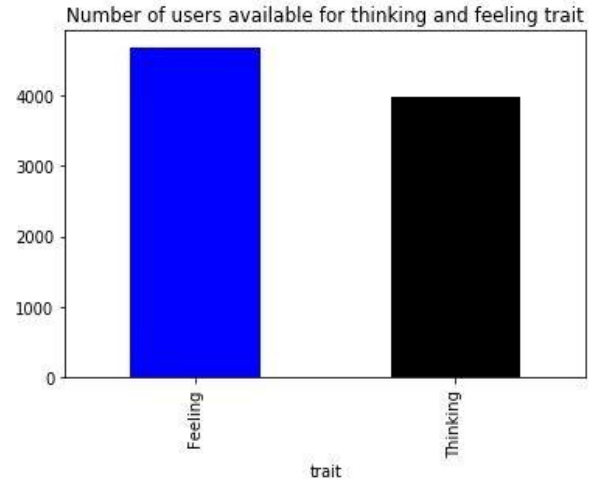

Fig. 3 User Count for Thinking VS Feeling Trait

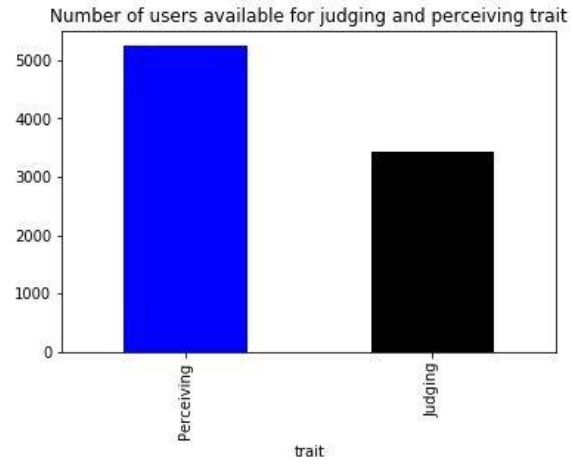

Fig. 4 User Count for Judging VS Perceiving Trait

IV. METHODOLOGY

To obtain accurate results while using NLP and SVM it is of utmost importance that steps like data preprocessing and feature extraction are carried out before the data is used by the intended algorithm. Since the original dataset was observed harboring imbalance between two pairs of traits, an under sampled version of the original dataset was created. This dataset contained 50 randomly selected posts from 50 users of each type, meaning 2,500 posts were present for each type, with the exception of types ESFP, ESFJ and ESTJ who had only 48, 42, 39 users respectively. The flow diagram for the methodology can be observed in fig. 5.

A. Data Preprocessing

It is a crucial step and involves the creation of a much more usable format from raw data. For our dataset we first checked if there were any null values, no null values were found. If the post created by a user contained the 4 coded MBTI type, then the MBTI type was removed, this was carried out to prevent inaccurate predictions. Next the post separator '| | |' was replaced by a semicolon. Thirdly all the https links that were present were replaced by the word 'link'. Since the same word in upper case and lower case would be treated differently all the uppercase words are transformed to lower case words. Tokenization, which refers to the process of generation of special words or phrases from a stream of data was performed. Finally word lemmatization was performed which changes the articulation of a word to its root form. Lemmatization was chosen over stemming as stemming does not consider the context of a word therefore words with different meanings will have the same base root.

B. Feature Extraction

Problems such as overfitting, redundant values and irrelevant data can lead to inaccurate predictions. Feature extraction is a way to combat the above difficulties. We employed "Term Frequency Inverse Document Frequency" (tf idf) which is a statistical approach used for determining the importance of a word in a document. Term frequency abbreviated as tf specifies how many times a given term occurs in a document. Inverse document frequency on the other hand evaluates the weightage of a word in reference to how common or rare the word is in the

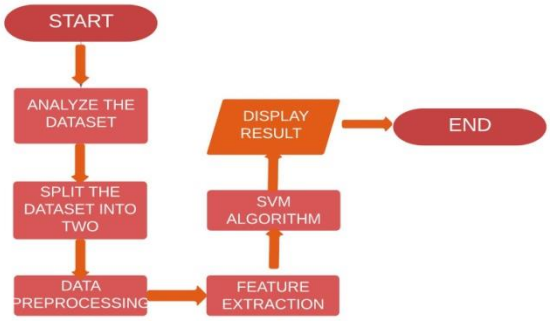

Fig. 5 Flow Diagram

document. It follows the notion that less common terms are more important frequent terms. Keeping this concept as its backbone tf-idf would then build a vocabulary of words with a unique integer number associated with each word. The maximum number of features that we inspected were 5,000.

C. Classifying Data

SVM was used to classify a person based on what kind of traits they possess. We have employed the linear kernel to analyze the accuracy percentage for the identification of each corresponding trait. The SVM algorithm was applied to all the trait pairs for the under sampled dataset whereas for the original dataset only two trait pairs namely thinking(T)/feeling(F) and judging(J)/perceiving(P) were analyzed through the SVM algorithm.

V. RESULT

Since only two trait pairs, thinking/feeling and judging/perceiving, had been balanced in the original dataset SVM had been applied only to them with the accuracy rates for those traits shown in table- II.

On the other hand, in the case of under sampled dataset all the corresponding traits were balanced so the SVM algorithm had been applied to all trait pairs. This is shown in table-III

Table-II: Accuracy Rates for Original Dataset

TRAIT	T/F	J/P
ACCURACY	78.5%	67.9%

Table-III: Accuracy Rates for Under Sampled Dataset

TRAIT	I/E	S/N	T/F	J/P
ACCURACY	58.9%	32.9%	30.7%	60.2%

It is observed that T/F has the highest accuracy rate for the original dataset and J/P for the undersampled dataset. On comparing the results obtained in table II and table III it is observed that the judging(J)/perceiving(P) trait has the most closely related accuracy rate meaning it is the most easily identifiable trait. In contrast the accuracy rates for thinking(T)/feeling(F) in both the tables is vastly different meaning this trait is the least easily identifiable trait.

VI. CONCLUSION AND FUTURE WORK

A person's writing style provides valuable information regarding the inner workings of their mind which can in turn help in the determination of their personality. The proposed model to achieve this goal is by classification of text using NLP and SVM which when performed on the MBTI dataset [11] produced the highest accuracy rate of 78.5% for the trait pair thinking/feeling. In the future the accuracy can be improved by considering a much larger and balanced dataset. The accuracy can also be improved by addition of more features.

ACKNOWLEDGMENT

A heartful thanks to Mrs. R.Angeline for her constant motivation and support throughout the tenure of this project. The development of this project would not have been possible without the Computer Science and Engineering department of our college SRM Institute of Science and Technology, Ramapuram.

REFERENCES

1. Michael Mesfin Tadesse,Hongfei Lin, and Bo Xu. Personality Predictions Based on User Behavior on the Facebook Social Media Platform. IEEE, 2018.
2. Tasfia Hoque, Raqeebir Rab,and Khushnoor Rafsan Jani Alam.Empirical Study on Personality Trait Classification by Food Related Preferences. In International Conference on Electrical ,Computer and Communication Engineering(EECE).2019.
3. Naveen.K.Kambham,Kevin.G.Stanley,and Scott Bell.Predicting Personality Traits Using Smartphone Sensor Data and App Usage Data.IEEE,2018
4. Guozhen Zhao,Yan Ge,Buying Shen,and Hao Wang.Emotion Analysis for Personality Inference from EEG Signals.IEEE,2017.
5. Anisha Yata,Prasanna Kante,T Sravani,and B Malathi.Personality Recognition using Multi-Label Classification.IRJET2018.
6. Tao Ding,Cheng Zhang, and Maarten Bos.Causal Feature Selection for Individual Characteristics Prediction. IEEE 30th International Conference on Tools with Artificial Intelligence.2018.
7. Junjie Chen,Zikai Wng,Qilian Liang.Bandwidth Allocation Based on Personality Traits on Smartphone Usage and Channel Condition.IEEE,2018.
8. Mihai Gavrilescu, and Nicolae Vizireanu. Predicting the Big Five personality traits from handwriting.EURASIP Journal on Image and Video Processing,2018.
9. Giulio Carducci, GiuseppeRizzo,Diego Monti,Enrico Palumbo, and Maurizio Morisio.TwitPersonality:Computing Personality Traits from Tweets Using Word Embeddings and Supervised Learning.2018.
10. Raad Bin Tareaf, Philip Berger,Patrick Hennig,and Christoph Meinel.Personality Exploration System on Online Social Networks:Facebook Brands As a Use Case.IEEE,2018.
11. MBTI) Myers-Briggs Personality Type Dataset.Version 1,2017.

AUTHORS PROFILE

Neha Gupta is currently pursuing her Bachelor degree in computer science and engineering from the prestigious SRM institute of science and technology. Her current interests lie in the domains of achine learning and data science. She is a meritorious student. She is planning to pursue her masters in one of the above domains.

Anirudh Madhavan is currently pursuing his Bachelor degree in computer science and engineering from the prestigious SRM institute of science and technology. His current interests lie in the domains of Algorithms and Machine Learning and is looking out for a prospective career in his domain. He is a meritorious student. He is planning to pursue his Masters in Machine learning.

Divya Duvvuri is currently pursuing her Bachelor degree in computer science and engineering from the prestigious SRM institute of science and technology. She is interested in the domain of machine learning and also intrigued in the domain of Mathematics. She is looking out for a promising career in her domain.

Angeline R is currently an Assistant professor at the computer science and engineering department of SRM institute of science and technology. Her field of expertise is Deep learning.