Predicting Myers-Briggs Type Index with Text Classification Using Data Mining Techniques

Soyadev Devadoss and Neelu Konkimalla
Department of Computer Science
Georgia State University
Atlanta, GA 30303

***Abstract*— Personality types often categorize people according to their tendencies to think and behave in a certain way. It is vital to sort people into meaningful groups, which makes people different. Myers-Briggs Type Index (MBTI) is designed to identify personality types, strengths, and preferences within the 16 different personality types. This study explored the use of machine learning to create a classifier that can sort people into their MBTI personality types based on social media posts. Social media is so widespread; such a classifier will have plenty of data to run personality tests, enabling more people to learn their MBTI personality style and produce accurate results.**

Keywords— Myers-Briggs Type Index, prediction, MBTI, post, social media, data mining

I. Introduction

Any personality assessment aims to enable individuals to explore their traits. To achieve personality type, social media posts have to play a very vital role. People tend to write differently, and they often select different subjects to discuss. In the Kaggle dataset from PersonalityCafe, an online forum dedicated to personality types. Furthermore, through understanding a person's personality traits, interests, and dislikes, social media platforms can improve the user experience. Personality is regarded as an influential yet imprecisely established construct in the empirical field of psychology. More practical, observational tests of current personality models would be beneficial to psychologists.

The Myers-Briggs Type Indicator is one of the models and the project aims to enhance understanding of it (MBTI). The use of machine learning to build a text-based classifier. For example, the use of social media posts as input and output to predict the author's MBTI personality type. Since the relationship between natural language and personality type is complex, creating an effective text-based classifier has significant potential consequences for psychology [4].

II. Background and Related Work

As a systematization of archetypal personality styles used in clinical practice, the MBTI personality classification system evolved from Jungian psychoanalytic psychology. There are 16 distinct personality types in the system, which are categorized into four binary orthogonal personality dimensions. It incorporates a four-letter code based on four axes, with each letter denoting the most prevalent characteristic on each axis.

The following are the indicators:

i. Extraversion (E) vs Introversion (I): a measurement towards determining whether an individual prefers their outer or inner world.

ii. Sensing (S) vs Intuition (N): a measurement of how much information is processed through the five senses versus perceptions by patterns by a person.

iii. Thinking (T) vs Feeling (F): a measurement of one's preference for rational values and reality overweighing other people's emotional viewpoints.

iv. Judging (J) vs Perceiving (P): a measurement of how much a person prefers a planned and structured life over one that is more flexible and spontaneous.

The predictive validity of the MBTI in terms of recurring personality characteristics is currently being debated. The Big Five personality classification system, in contrast to the MBTI system, is the most widely used personality type system in Psychometrics. Extraversion, agreeableness, openness, conscientiousness, and neuroticism are the five statistically orthogonal personality dimensions measured by this personality system.

In comparison to the MBTI personality type system, the Big Five personality type system is statistically derived to have predictive power over observable features in an individual's life, such as income, education level, and marital status, and it is consistent throughout a person's lifetime. However, work by J. W. Pennebaker and L. A. King [3].

There are important similarities between four of the Big Five personality characteristics and the four MBTI dimensions: Agreeableness with Thinking/Feeling, Conscientiousness with Judging/Perceiving, Extraversion with Extraversion/Introversion, and Openness with Sensing/Intuition. These similarities justify an attempt to model the relationship between writing style and recurring personality traits in the sense of our project and the MBTI personality system's popularity. There is a lack of current research on predicting MBTI personality types from textual data. Nonetheless, both machine learning and neuroscience have made significant progress. Jonathan S. Adelstein [2] discovered the neural correlates in his research.

III. Proposed Approach

The goal of this project is to predict the MBTI personality of a person. A couple of model are used to predict the text of the post. A type analysis is conducted to predict the characteristics among the post from social media.

Dataset Description

For training, the model uses forum posts from PersonalityCafe.com. This dataset has already been made available on Kaggle [1], but it is not a piece for competition;

it is a dataset to analyze with. There are 8,675 respondents in this research. There are two columns in this section: one being the type and the other being the collection of posts made by the person of the type.

There are a total of 16 MBTI codes [2]. The following are the 16 codes:

ISTJ: The Inspector (Introverted, Sensing, Thinking, Judging)

ISTP: The Crafter (Introverted, Sensing, Thinking, Perceiving)

ISFJ: The Protector (Introverted, Sensing, Feeling, Judging)

ISFP: The Artist (Introverted, Sensing, Feeling, Perceiving)

INFJ: The Advocate (Introverted, Intuitive, Feeling, Judging)

INFP: The Mediator (Introverted, Intuitive, Feeling, Perceiving)

INTJ: The Architect, (Introverted, Intuitive, Thinking, Judging)

INTP: The Thinker (Introverted, Intuitive, Thinking, Perceiving)

ESTP: The Persuader (Extraverted, Sensing, Thinking, Perceiving)

ESTJ: The Director (Extraverted, Sensing, Thinking, Judging)

ESFP: The Performer (Extraverted, Sensing, Feeling, Perceiving)

ESFJ: The Caregiver (Extraverted, Sensing, Feeling, Judging)

ENFP: The Champion (Extraverted, Intuitive, Feeling, Perceiving)

ENFJ: The Giver (Extraverted, Intuitive, Feeling, Judging)

ENTP: The Debater (Extroverted, Intuitive, Thinking, Perceiving)

ENTJ: The Commander (Extraverted, Intuitive, Thinking, Judging)

Pre-Processing

Looking at other machine learning studies of MBTI, the researchers rarely cleaned their data sets to match the real proportions of MBTI forms in the general population [1]. Some cleaning of the proportional representation of each MBTI sort would be required because the raw data set is severely disproportional relative to the approximately uniform distribution for the general population. To avoid any misinterpretation of results due to distorted representation of classes in the test set, we artificially made our test set to reflect the proportions found for each form in the general population.

The process includes the number of people in the dataset by MBTI sort is compared to the global population percentage representation of each type mapped onto the graph in Fig 1.

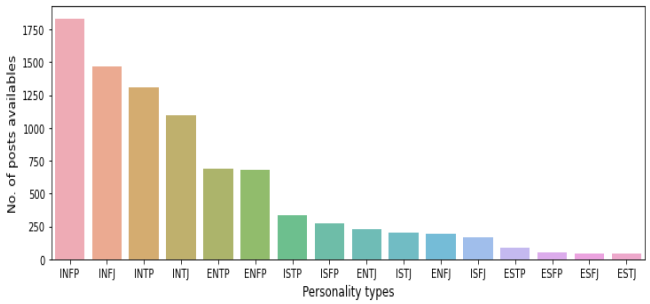

Fig. 1. MBTI types are defined in a non-uniform way in this data set.

The dataset is displacing unbalanced groups. Certain personality styles have a lot more data than others like with *INFP* being the most popular (Introvert Intuition Feeling Perceiving). However, inferring that people who regularly comment on social media are more introverted, perceptive, and emotional. Having unevenly distributed data can be challenging as it does not present the 16 personality types evenly.

As the data set comes from an Internet forum, some words can be removed to make the values meaningful as possible. Some data points contain links to websites in the form of videos or images. This potentially does not add to the dataset as it is only focusing on the English language. Further, based on the links, it will not provide any information about an individual's personality. Also, stop-words and stem-words can be removed as most of them do not give any helpful information. Few examples of stop words include "I," "to," "the," etc.

Lemmatization allows grouping together the different inflected forms of a word to analyze as a single item. For example, the words "rocks" and "rock" have the same base or dictionary form of a word, and it will be considered to one single form.

Lemmatization is beneficial as it relies on correct language data to identify a word and give accurate results. Lemmatizations are the process of joining words that have similar meanings into a single word. Lemmatization examples include: better: fine, corpora: corpus, rocks: rock, corpora: corpus

Word clouds are a data visualization technique that contains words to represent their frequency. If a word is large, then the word is common among that personality type. Below is the picture that highlights the common words within the given type. This cluster is a powerful tool that helps to see the correlation of specific words in the database.

Fig. 2. Number of irrelevant words present in the dataset

We can see that the dataset contains a large number of obsolete terms (e.g. ha, ar, Ti, etc.) that will need to be deleted. Interestingly, the names of MBTI personality types themselves are among the most common terms in word clouds of individual personality types shown in Fig2. As a result, as part of our pre-processing level, we'll need to clean up our posts by removing these MBTI terms from each of them before training the model for better evaluation performance.

IV. FEATURE ENGINEERING

Splitting into X and Y feature

Label Encoder

Sklearn library provides a function that transforms the levels of categorical features (labels) into a numeric form so that it can be machine-readable. Labels with a value between 0 and n classes 1, where n is the number of distinct labels, are encoded. When a mark is repeated, it is given the same value as before[8]. Using Label Encoding to encode personality type, convert MBTI personality (or target or Y feature) into numerical form.

	type	posts	no. of. words	type of encoding
0	INFJ	enfp intj moments sportscenter plays...	430	8
1	ENTP	finding lack these posts very alarming eo...	803	3
2	INTP	good course which know thats bles...	253	11
3	INTJ	dear intp enjoyed conversation other eos...	777	10
4	ENTJ	youre fired eostokendot thats another silly...	402	2
5	INTJ	eostokendot science perfect eostokendo...	245	10
6	INFJ	cant draw nails haha eostokendot those w...	970	8
7	INTJ	tend build collection things desktop th...	140	10
8	INFJ	sure thats good question eostokendot dist...	522	8
9	INTP	this position where have actually pe...	130	11
10	INFJ	time parents were fighting over dads affair...	1072	8

Fig. 3. Lable Encoder present in the dataset

Label encoding is used over one-hot encoding to save pre-processing time, and MBTI has predefined 16 values, and assigning unique integers based on alphabetical ordering appears to be a viable choice as shown in Fig 3.

Count Vectorizer

Count Vectorizer is a program that converts a collection of text documents into a vector of term/token counts and builds a vocabulary of known terms, as well as encoding new documents with that vocabulary. It also allows text data to be pre-processed before being converted into a vector representation. Since Count Vectorizer only counts the occurrences of each word in its vocabulary, incredibly common terms like 'the,' 'and,' and so on would become very important features while adding little meaning to the text if stop words='English' is used. This is a vital step in pre-processing because not taking certain terms into account can also strengthen our model. As a result, our dataset now has 98555 features for 8466 rows (users).

V. MODELS

Train the model using a variety of machine learning algorithms, such as Random Forest, XGBoost to find the best classifier.

Algorithms

One popular technique, the *Random Forest* algorithm, can perform both regression and classification tasks. It is built from decision trees that are easy to use and interpret. The random forests are created with many decision trees. More trees in the forest mean the more robust the prediction resulting in higher accuracy. In a random forest, multiple trees will grow, and each tree will give a classification[5]. The forest chooses the classification based on the most votes.

XGBoost algorithm is a framework that can run on multiple languages like Java, Python, C++, etc. It is also portable, making it platform-free and can run on mac, windows, and Linux, etc. Further, it is all in one when it comes to small to medium structure datasets and performs well in classification, regression, ranking, and other problems. The basic idea of the algorithm is when multiple trees are building on top of each other to correct the errors of the previous tree before[7]. This simple idea is straightforward to use, computationally efficient and makes an accurate prediction.

Training and Evaluation

The algorithm is split into a 60-40 ratio where 60% is dedicated to training the model and 40% is dedicated to testing the algorithm. In this way, one can compare the predictions from the model with the testing data. The training set must be separate from the test set. The training side has the actual value that the model should have predicted. And testing data is a part of the same dataset but it was never shown to the model before. Overall, this helps pay attention to minor details and optimizes the dataset accuracy. For the random forest classifier, the accuracy is 39.53% and the XGboost classifier is 57.87%. This process was repeated for a 70-30 ratio as well as splitting the data into training and testing sets. It is performed multiple times to prevent overfitting.

HeatMap for determining the correlation between the character types

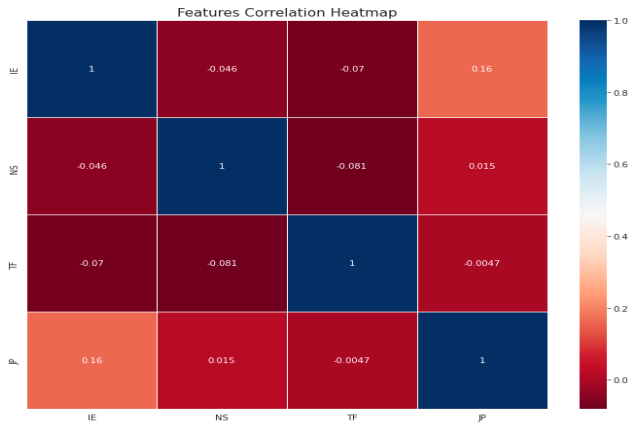

Fig. 4. Correlation Between Different Personality Types

The simulation result in Fig 4 shows the prediction of the various personality types. Observing the association heatmap findings for the various personality styles shows the relationship that exists. Between I/E and J/P, there is a correlation of 0.16. This suggests that their characteristics appear to converge. People's reactions and interactions with the environment around them are defined by I/E. J/P is a term that explains how people communicate with the outside world, and how they interact with other people.

Four Axis Classifiers across MBTI axis

There were 8675 rows and 2 columns in the dataset. If an individual has the letters I, N, T, and J, the value across the four MBTI axes, IE, NS, TF, and JP, will be 1, otherwise 0. This will allow us to determine how many Introvert posts vs. Extrovert posts are present in the dataset. This is achieved to search the dataset for all of the MBTI Personality Indices individually.

	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. 5. Four Axis Classifiers across MBTI in the dataset

This Fig 5 will allow us to determine, for example, how many Introverts vs. Extrovert posts are present in our labeled Kaggle dataset out of all the given entries. This is done to explore the dataset for all of the MBTI Personality Indices individually.

$$Count = No.of\ posts\ in\ one\ class / Total\ no.of\ posts\ in\ the\ other\ class$$

```
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. 6. Unequal Distribution Four Axis Classifiers

Fig 6 deduces that there is unequal distribution around each of the four axes in our dataset's entries. IE: E is the majority, while NS:S is the majority. TF and JP, on the other hand, have a lot less of a difference between them

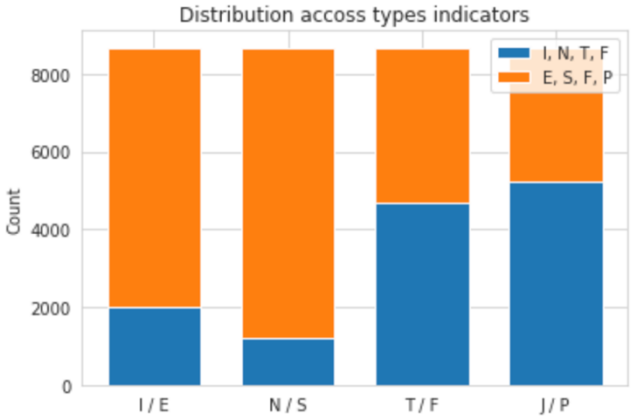

Fig. 7. Type indicators

In Fig 7 displayed distribution across each letterform is assumed to be independent of the others in our model, i.e., a person's introversion/extroversion is unrelated to their judgment/perception. Nonetheless, users can put them to the test using a heat map below.

Tf–Idf (term frequency-inverse document frequency) for feature engineering assesses the relevance/importance of a word to a document in a corpus of documents. It is very useful for scoring words in machine learning algorithms for Natural Language Processing since we train individual classifiers here.

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

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

Fig 8: Number of irrelevant words present in the dataset

In Fig 8 count vectorizer and of Tf-IDF vectorizer is used to vectorize our model, holding the terms appearing between 10% and 70% of the posts.

Model Accuracy

The Random Forest Algorithm and the XGBoost Algorithm are used to train each MBTI personality form individually, and the results are as follows:

Random Forest Classifier Accuracy in Fig 8

```
IE: Introversion (I) / Extroversion (E) Accuracy: 77.33%
NS: Intuition (N) / Sensing (S) Accuracy: 86.03%
FT: Feeling (F) / Thinking (T) Accuracy: 67.66%
JP: Judging (J) / Perceiving (P) Accuracy: 62.80%
```

Fig. 8. Random Forest Classifier Accuracy

XGBoost Accuracy Fig 9

```
IE: Introversion (I) / Extroversion (E) Accuracy: 77.58%
NS: Intuition (N) / Sensing (S) Accuracy: 86.03%
FT: Feeling (F) / Thinking (T) Accuracy: 71.22%
JP: Judging (J) / Perceiving (P) Accuracy: 62.49%
```

Fig. 9. XGBoost Accuracy

VI. PERSONALITY PREDICTION

We have tested the algorithm to recognize the strengths of a person that can be helpful. The post that we made is

"*Hi. I'm Neelu. I'm currently a Junior at Georgia State University. I love having coffee while watching videos on YouTube. I'm a dog person and enjoy playing Frisbee with it*."

INFJ's result means that an individual is gentle, caring, complex, and highly intuitive. The MBTI can provide insight without taking formal quizzes. One can determine some of the tendencies with just a social media post.

VII. CONCLUSION

We have trained two Machine Learning Algorithms – Random Forest Classifier and XGBoost Classifier. The initial accuracy of the models was low when the models considered only the personality type. Then, we improved the model by taking individual personality typed such as I/E, N/S, T/F, J/P and predicted the model accuracy for each character type and the model accuracy was very high compared to the previous models. The dataset obtained is from Kaggle which contains the uneven distribution of data which makes the model biased over a particular personality type. The project can be further improved by obtaining data directly from Twitter or Facebook using API calls and analyzing that data for a particular personality type.

REFERENCES

[1] Mbti Kaggle data set. https://www.kaggle.com/datasnaek/mbti-type.

[2] Estp archetype. https://www.16personalities.com/estp-personality

[3] Jonathan S. Adelstein, Zarrar Shehzad, Maarten Mennes, Colin G. DeYoung, Xi-Nian Zuo, Clare Kelly, Daniel S. Margulies, Aaron Bloomfield, Jeremy R. Gray, F. Xavier Castellanos, and Michael P. Milham. Personality is reflected in the brain's intrinsic functional architecture. PLOS ONE, 6(11):1–12, 11 2011.

[4] James W Pennebaker and Laura A King. Linguistic styles: Language use as an individual difference. Journal of personality and social psychology, 77(6):1296, 1999.

[5] K. R. Scherer. Personality markers in speech. Cambridge University Press., 1979.

[6] Mayuri P. Kalghatgi, Manjula Ramannavar, and Dr. Nandini S. Sidnal. A neural network approach to personality prediction based on the big-five model. International Journal of Innovative Research in Advanced Engineering, 2015.

[7] M Komisin and Curry Guinn. Identifying personality types using document classification methods. Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS-25, pages 232–237, 01 2012.

[8] Gjurkovic, M.; Snajder, J. Reddit: A gold mine for personality prediction. In Proceedings of the Second Workshop on Computational Modelling of People's Opinions, Personality and Emotions in Social Media, New Orleans, LA, USA, 6 June 2018; pp. 87–97. Available online: https://peopleswksh.github.io/pdf/

[9] Myers, I.B.; McCaulley, M. Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator, 15th ed.; Consulting Psychologists Press: Santa Clara, CA, USA, 1989.