Define people's MBTI personality through Machine learning

Abstract

Personality tests can help us better understand ourselves and provide guidance on career paths. The motivation for building such a classifier is that companies and organizations would like to provide more customized products or services. It is valuable for them to know their customers' MBTI types and provide suitable services. However, nowadays the MBTI test takes quite a long time to complete. Normally people will lose patience in the end and randomly select the answer just to complete the test, leading to potential inaccurate test results. For companies, it is difficult and ineffective to collect customers' personality types by doing the test. Moreover, some customers may not be willing to share their MBTI type with the company due to privacy concerns. To save time and effort spent on the tests, we propose a machine learning solution, identifying people's MBTI personality through their posts in the forum. We explored two ways of model building, multiclass classifiers and combinations of binary classifiers, compared their performance, and discussed the potential limitations and business values. The GitHub link for this project is https://github.com/lychen99/BT5153-MBTI-Group16

1. Introduction

1.1 Background

Personality is the characteristic sets of behaviors, cognitions, and emotional patterns that evolve from biological and environmental factors. For years, people have been trying to link individual behavior with their personality.

The Myers–Briggs Type Indicator (MBTI) is one of most popular personality tests in the world, which is used in businesses, online, for fun, for research and lots more. It attempts to assign people to four categories:

- introversion(I) or extraversion(E)
- sensing(S) or intuition(N)
- thinking(T) or feeling(F)
- judging(J) or perceiving(P)

1.2 Problem Statement

A typical MBTI questionnaire contains about 100 multiple choice questions, which usually takes about 30 to 60 minutes to complete. Nowadays, people are getting more and more impatient with filling out questionnaires. People probably won't be willing to take time to test for MBTI results, or they are likely to randomly choose an answer in order to quickly finish the test.

In the meantime, companies and organizations want to provide highly personalized products and services for their customers to achieve better experience, while the application of personality in customized services is not extensive due to difficulty to collect this data. The objective of this project is to solve these problems by designing a machine learning system that evaluates individuals' MBTI results by analyzing their recent posts on the Internet rather than through questionnaires.

1.3 Objective

We aim to apply text analytics techniques to understand individual Myers-Briggs Personality Type (MBTI) from the post which user posted in the social media, so that we can:

- Identify their MBTI personalities type
- Precisely push relevant products and services for people with different personalities for social media marketing purposes.
- Increase pairing success rate by recommend people with similar personalities for dating applications and websites
- Increase the collaboration between different personalities, in hope to increase workplace diversity

2. Data and Preprocessing

2.1 Data Description

Our dataset is obtained from Myers-Briggs Personality Type Dataset from Kaggle (Mitchell, 2018). According to Mitchell, the data was collected from the Personality Cafe Forum. Personality Cafe Forum is a popular forum community where people with all ranges of personality

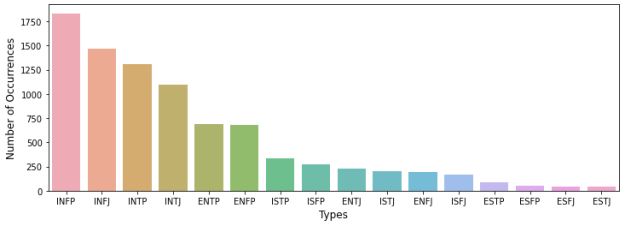
types post and discuss interests, health, behavior, care, personality and more.

The dataset contains over 8675 rows. Each row is data for a person, and it records the type of the individual and this individual's posts.

Table 1. Dataset Description

Variable	Data Type	Example
Type	String	INTJ
Posts	String	'Dear INTP, I enjoyed our conversation the other day…

The data is highly imbalanced, with more posts from top 6 personality types, consisting of 81% of total posts. The rest 10 personality types have less number of posts (19% of total posts).



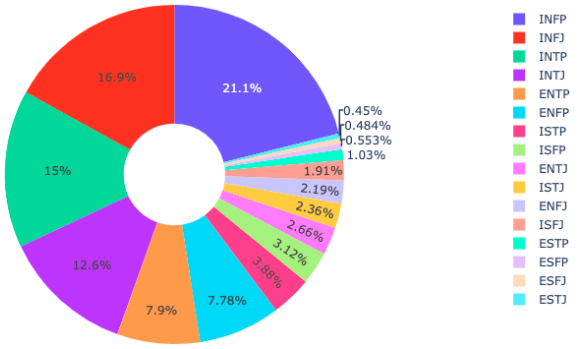


Fig 1. Distribution of 16 personalities

We want to dive into each of the 4 different scales, to see which scales have more imbalanced data distribution. It is observed that Introvert/Extrovert, Sensing/Intuition scales have more imbalanced distribution, specifically, we have less data from 'Extravert' and 'Sensing' types. Personal attributes distribution in Thinking/Feeling, Judging/Perceiving scales are relatively balanced. This is probably because 'Introvert' and 'Intuition' types of individuals like to post online.

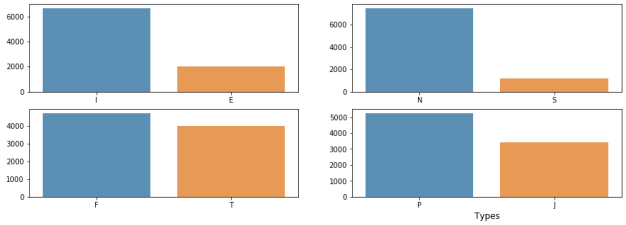


Fig 2. Distribution of each of the 4 scales (E/I, S/N, T/F, J/P)

2.2 Data Preprocessing

Since both traditional machine learning models (logistic regression, random forest, etc.) and deep learning models (CNN and BERT) will be examined in the following part, different kinds of data preprocessing methods are required. For the TF-IDF algorithm, heavy cleaning should be done to our dataset, while the BERT model may give a better performance along with light cleaning (Alexander, 2021). In this regard, we conducted a two-step data preprocessing. In the first step, we only converted the text to lowercase and removed url, digits, punctuations, diacritics, and whitespace. In the second step, we further cleaned our dataset by removing the stop words and performing lemmatization. Figure 3 shows the detailed data preprocessing jobs.

Fig 3. Data preprocessing

3. Exploratory Data Analysis

Before deep diving into models, we want to have an overall understanding of our data and text content, hence, several EDA methods are performed.

3.1 Length of Posts

We calculated the length of each post and found that the average word count is 614 per post. The distribution across different personality types is similar. ESFP type has slightly shorter average post length, but it could be due to smaller data samples for this type.

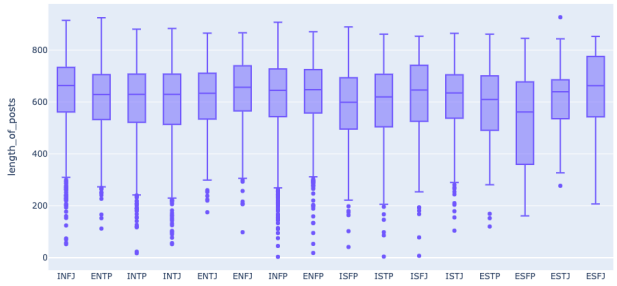


Fig 4. The length of posts for all personalities

3.2 Word Cloud

A word cloud is a visual representation of text data. The more important or frequently occurring words are shown with larger font size.

We generated word clouds for each type of personality, with max 30 words in a graph. To avoid noise from common words across different types, the original top common words such as 'people', 'think', 'know' are removed.

We selected some types to show in the figure below. The insights are

- The posts normally include the personality type of the individual
- Word *'feel'* occurs frequently for types with F (feeler). This can be understood as people with the Feeling (F) trait follow their hearts and emotions
- Word *'friend'* is more common for 'Extrovert'. it is reasonable as extraverts are interested in engaging with their environment, including the people around them.

Fig 5. Word cloud analysis

3.3 Sentiment Analysis

Sentiment analysis is the practice of using algorithms to classify text into overall positive or negative categories. We want to leverage sentiment analysis to understand whether different personality types tend to show more positive or negative emotions in their posts. We performed sentiment analysis using the NLTK library. The output is a dictionary of different scores: negative, neutral, and positive scores, which are then used to calculate the compound score. The compound scores can range from -1 to 1, and are visualized using Altair in the figure 6 below.

It clearly shows that the majority of posts are positive. Certain types have slightly higher percent of negative posts, such as ESTJ, INTP, ISTP. One explanation for ESTJ having more negative posts is that they are generally more judgemental, stubborn, and easily agitated, thus may have more negative or critical statements in their posts.

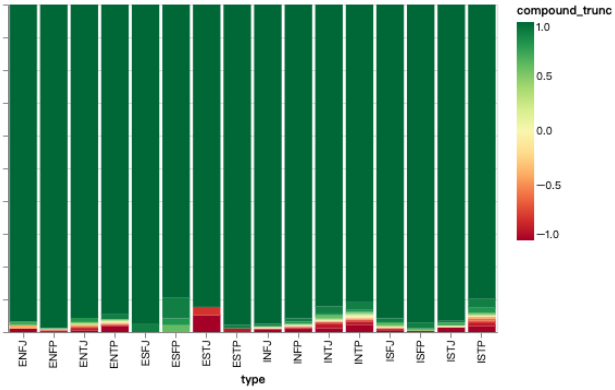


Fig 6. Sentiment scores for all personalities

4. Method 1:

We constructed two types of models to predict one's MBTI type, one is a multi-class classifier while the other one is a binary classifier. In this part, we will mainly show the result of multi-class classifiers, including TF-IDF, Doc2Vec, and neural network techniques. Details about the binary classifier will be covered in Method 2 part.

4.1 TF-IDF

We implemented TF-IDF to transform the textual data to a vector, and then we run several classifiers based on this TF-IDF vector. Since our dataset is highly imbalanced, both accuracy and f1-score are regarded as evaluation metrics. The result is listed in Table 2, and it turned out that XGBoost outperformed all other models in terms of accuracy and f1-score.

Table 2. Model Performance with TF-IDF

Model	Accuracy	F1-score
Logistic Regression	0.6340	0.6019
Random Forest	0.3844	0.3115
XGBoost	**0.6801**	**0.6701**
LightGBM	0.6651	0.6509
Multinomial Naive Bayes	0.2126	0.0770
LinearSVC	0.6674	0.6582

4.2 Doc2Vec

Though TF-IDF is one of the most popular techniques utilized in text classification, it doesn't take the word order into account. To further explore the semantics and syntactic order of words in a complex text, we decided to apply Doc2Vec to our dataset, which is an unsupervised algorithm based on word2vec method that can generate "a numeric representation of a document (Shperber, 2019, p.1)". After we constructed the feature vector, we trained several classifiers, including Random Forest, XGBoost, and etc., to examine whether the document embedding model could improve the model performance. The result is listed in Table 3.

Table 3. Model performance with Doc2Vec

Model	Accuracy	F1-score
Logistic Regression	**0.4144**	**0.3584**
Random Forest	0.2691	0.1786
XGBoost	0.3383	0.2632
LightGBM	0.3538	0.2747
LinearSVC	0.3786	0.3150

The result shows that logistic regression outperforms its counterparts in terms of accuracy and f1-score, and its confusion matrix is shown in Figure 7. The confusion matrix indicated that even if the model failed to correctly predict the MBTI type of a specific post, it can still identify the post as another MBTI type which generally shares 2 or more categories in common with the real type, suggesting that people with similar personalities may have similar texting habits.

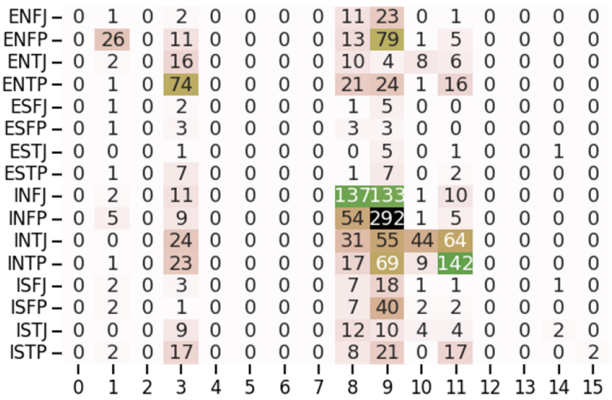


Fig 7. Confusion matrix of logistic regression model

4.3 Neural Network

We tried two types of neural network models to build the multi-class classifier. We split the dataset into a training set (80%), a validation set (10%) and a test set (10%). Training set is used for model training, and the validation set and test set are used for model performance evaluation.

4.3.1 CNN (Convolutional Neural Network)

For CNN, we set the size of vocabulary to 1000, embedding dimensions to 20, dropout rate to 0.2, epoch to 10 and batch size to 64. The optimizer we leverage on is Adam. We use a combination of 2-grams, 3-grams and 4-grams filters and max pooling method. We use dropout, learning rate decay and early stopping to prevent overfitting. The initial learning rate is 1e-2. Decay steps are 12000. Decay rate is 0.8.

The performance on the test set of CNN is presented in Table 4.

Table 4. CNN Prediction Performance on Test Set

	Precision	Recall	F1-score
Accuracy	0.68		
Macro avg	0.61	0.50	0.52
Weighted avg	0.68	0.68	0.67

4.3.2 BERT (Bidirectional Encoder Representations from Transformers)

For the BERT classifier, we used a pre-trained BERT layer combined with a dropout layer. The optimizer we choose is Adam. The batch size is 8. The dropout rate is 0.5 and we use a ReduceLROnPlateau learning rate scheduler. The initial learning rate is 1e-5.

The performance on the test set of BERT is presented in Table 5.

Table 5. BERT Prediction Performance on Test Set.

	Precision	Recall	F1-score
Accuracy	0.63		
Macro avg	0.49	0.48	0.48
Weighted avg	0.62	0.63	0.62

5. Method 2: Binary Classifier

For method 2, our idea is to based on MBTI 4 measurement dimensions, we train 4 classifiers individually to classify their personalities

The Myers Briggs Type Indicator (MBTI) divides everyone into 16 distinct personality types across 4 axis:

Introversion (I) – Extroversion (E)

Intuition (N) – Sensing (S)

Thinking (T) – Feeling (F)

Judging (J) – Perceiving (P)

5.1 Algorithm used

Logistic regression, KNN, SVM and XGBoost are applied to classify each axis. MBTI type indicators were trained individually, and the data was split into training and testing dataset using the train_test_split() function from sklearn library. Totally, 70% of data was used as the training set and 30% of the data was used as the test set. The model was fitted onto the training data and the predictions were made for the testing data.

5.2 Results

The results of all the algorithm were shown in the below table:

Table 6. Results of different algorithms in 4 axis

	Extraversion(E) - Introversion(I)	Sensing (S) - Intuition (N)	Thinking(T) - Feeling(F)	Judging (J) - Perceiving(P)
Logistic Regression	0.7754	0.8606	0.7244	0.6451
KNN	0.7667	0.8582	0.5470	0.4020
SVM	0.7796	0.8603	0.7262	0.6587
XGboost	0.7737	0.8585	0.7028	0.6437

After calculating the average accuracy of four dimensions, we select SVM to do the evaluation of method 2 because it has the best average accuracy.

5.3 Evaluation of Method 2

In order to evaluate method 2 final prediction accuracy, we randomly select 100 data from the dataset, input to the model after preprocess and vectorizing.

The result of the 100 data prediction were shown below:

Table 7. Method 2 prediction result

Correct Prediction	Wrong Prediction		
	1D Wrong	2D Wrong	3D Wrong
20	50	26	4

Accuracy of correct prediction: 20/100 = 20%

% of one dimension prediction wrong: 50/100 = 50%

% of two dimension prediction wrong: 50/100 = 26%

% of three dimension prediction wrong: 50/100 = 4%

From the result we found that the accuracy of method 2 is lower than method 1. We think there may be other factors affecting the accuracy instead of model building issues.

5.4 Discussion

Though "the administration and interpretation of the MBTI is a huge business and force in shaping the general public's perceptions of psychology (Stein and Swan, 2019, p.1)," arguments and criticisms of the MBTI never stop. First, some studies suggested that the MBTI test has poor reliability and poor validity. To be more specific, poor reliability means that the test results can vary a lot when the same individual retaking the test (Sambursky, 2022), just like Pittenger (2005, p.214) pointed out, "Across a 5-week re-test period, 50% of the participants received a different classification on one or more of the (MBTI) scales." As for MBTI's poor validity, Boyle (1995, p.73) argued that "since MBTI types are not source traits verified factor analytically, predictions based on these surface traits are inevitably less powerful and remain somewhat speculative." Also, some studies believed that MBTI "is not comprehensive because its categories do not capture the full extent of personality (Sambursky, 2022, p.1)."

Most importantly, MBTI overlooked the fact that "personality traits are not static (Sambursky, 2022, p.1)." For instance, MBTI assumes a person is either an Extrovert or an Introvert, however, the distribution of personality traits may be a bimodal distribution rather than a normal one, suggesting that "personality dimensions are continuous, with persons being more or less extraverted or introverted (Riggo, 2014, p.1)." And that can partially explain why our model failed to meet expectations, because there isn't any clear boundary between each personality type, for instance, even if a person claimed himself to be an INTJ, he can also have characteristics of an Extrovert or a Feeler.

5.5 Interpretability

Regardless of the high accuracy of our model, it is currently still a black-box and hard to interpret. We need to add a model interpretation for people to comprehend and trust it.

SHAP is used to interpret how our model makes predictions on a global and local scale.

Taking the SVM classifier–Extraversion(E) v.s. Introversion(I) as an example, we randomly take 100 posts from training data and explain how each word impacts the output of the model. On the global scale, the impact of each word is stacked to create the importance plot, as shown in figure 8.

The model has 2 lables: lable 0 stands for Introvert, and lable 1 is Extrovert. The top most important words as calculated by SHAP are "ne" or intuition and extraversion, "awsome", "fun", "love", ect. These are all the words which we naturally relate to Extraversion, suggesting that our model has managed to learn the important features from posts, and does have high interpretability.

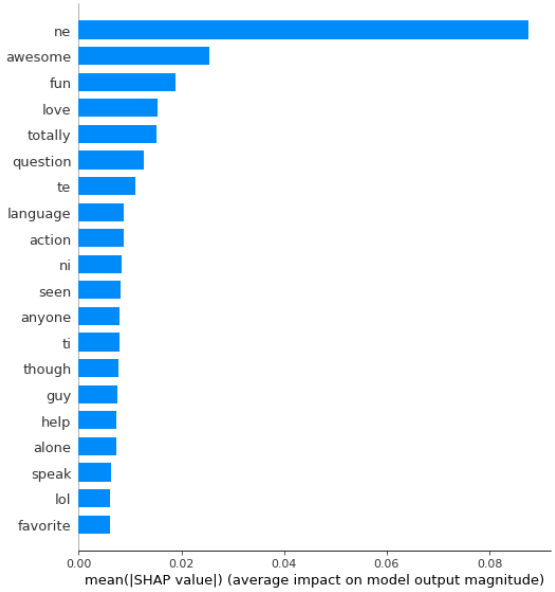


Fig 8. SHAP summary plot for Extrovert / Introvert

SHAP is also able to interpret our model locally. The following figure shows a Force Plot of an individual post, and explains how each word contributed to that post's prediction. Let us take this clean posts as an example:

"cal definitely without doubt nt search answer way scream rational think observant think action acting n awesome article one question could totally wrong

pictured roo ne piglet ni owl ti mostly roo spontaneous look proud trying think know irl u"

Here, the predicted label is 1, or Extrovert, whereas the true label is also 1. The base value is the average of the model output over the training dataset. In this case, the training dataset is the 100 posts which we randomly sampled, and the base value = 0.3. The numbers on the plot are the value of the feature for this post. Red arrows represent features that pushed the model score higher, or more Extroverted, and blue represent features that pushed the score lower, or more Introverted. The bigger the arrow, the bigger the impact.

Fig 9. SHAP force plot for a specific post

For the most part, the red arrows are related to Extraversion, like the words "awesome" and "action". But there are few counter-intuitive ones. The word "fun" is interpreted as a contribution to Introversion score.

To conclude, SHAP can help us understand how each word collectively contributes to model prediction, although it does have some errors. Nevertheless, it provides a useful tool to open the black-box and interpret our model.

Comparison between Method 1 and Method 2

Analysis of accuracy, method 1 has better performance than method 2. For some models such as XGBoost and CNN, they can reach 0.68 which is a very good accuracy. However, for method 2, the accuracy of 100% correct prediction of 4 dimensions is only 0.2. which seems to indicate a weak overall ability of our model to correctly classify all four MBTI dimensions.

Even though method 1 may achieve higher accuracy of perfect classification, they do have a risk of getting their prediction completely wrong. Method 1 treats all classes as independent of each other, so fails to capture the in-built relatedness of some types to other types. For example, INFP is much more similar to INTJ than it is to ESTJ. For method 2, achieve lower rates of perfect classification in exchange for higher rates of approximately correct classification.

6. Business Potential Area

Our model is expected to accurately predict MBTI and will bring business value in almost anything which involves people (Peter, 1997). We will expand on 4 of the potential model applications.

Increase successful date matching rates for Dating apps/websites: Our work will improve pairing success rates for dating applications and websites. Through our work, the pairing will become much more effective. It will greatly increase the pairing success rates and bring a good reputation for the company.

Product recommendations and personalized marketing: Customers are increasingly expecting companies to treat them as individuals, rather than mass marketing (Lindecrantz et al., 2020). Personalization can also improve customer retention and brand affinity. Understanding how different personality types behave can help companies to understand consumer preferences, how to reach them, and their acceptance to a marketing campaign (Evans, Madeline, 2021). Knowing the audience and personalizing products can boost engagement, brand reputation and loyalty. The need for segmentation presents a promising application to our model.

Utilize the power of diversity in the workplace: Another potential application builds on top of the buzzword "diversity". In 2015, McKinsey reports that those in the top quartile for ethnic and racial diversity in management were 35% more likely to have financial returns above their industry mean (Hunt et al., 2021). In addition, knowing people's personalities and leveraging it could help to build stronger, more effective teams in the workplace. Like design various team-building exercises for teams with different backgrounds, or design individualized training programs. Our model is expected to accurately identify an individual's personality type based on one's posts, which is less subjective compared to the self-reported values. Therefore, the model provides one more facade to consider when making crucial executive decisions or executive talent management, reducing the risk of homogeneity in the workplace.

Recommendation for job matching websites: Our model is useful for organizations that provide platforms to link talents and job positions. It adds one more dimension to consider when pushing relevant jobs to people with different MBTI. Organizations like JobsDB and Linkedin aim to tap into potential customers and push suitable jobs to them. Our work will contribute some value to that. As we all know, different personalities have different expectations of work and personal development, for example people with INTJ type would be interested to lead or find management level positions in the job market,

They may have a high interest in positions that require leadership skills. Knowing people's personality could help websites like linkedin to push accurate recommendations.

7. Future Work

Although the model is able to achieve reasonable accuracy, we have identified the following two areas of improvement.

The current datasource could be further improved on two aspects. First is to avoid dataset homogeneity, as the current dataset is only collected from a single source, the Personality Cafe Forum, and is therefore prone to the risk of overfitting to that datasource. We could reduce bias by including more posts/comments from various other data sources, and increase the model's ability to generalize. Another aspect is to have a more balanced class distribution. As we mentioned in the data preprocessing section, the personality distribution is highly imbalanced. We could get additional data for all personalities and enable the model to learn equally from each of them.

In addition, the model should be robust enough to handle uncertainties in different situations. For example, the obtained information or the data available for analysis in real-life could look quite different from the posts we used in this project. The model needs to be generalized to learn from various kinds of data, not only posts, while achieving similar performance at the same time.

8. Conclusion

In this project, we used 2 methods to identify a person's MBTI type. Method 1 is to classify an instance into one of the 16 categories, and is able to achieve a reasonable F-1 score of 0.67. However, Method 1 is not without its limitations, as it assumes a person can only fall into one of the 16 personality types, and that the MBTI types are independent from each other. To address these limitations, we have designed Method 2, which build 4 binary classification models, with one classifier for each dimension(Introversion (I) – Extroversion (E); Intuition (N) – Sensing (S); Thinking (T) – Feeling (F); Judging (J) – Perceiving (P)), and combine these 4 results to get a final prediction. Method 2 has a lower combined accuracy in terms of correct prediction across all 4 dimensions, but it has a different prediction objective: Method 2 focuses more on binary classification in each dimension, not how well it can label an individual into one of the 16 categories.

Since the accuracy in each individual dimension is quite high (average accuracy is 0.76), Method 2 could be more useful in situations where knowing which of the 16 MBTI types a person has is less important; and just knowing one personality dimension is enough to satisfy the application's requirement. For example, in companies management recruitment, they may care more to know whether a candidate is of type Judging(J) or Perceiving(P), not which of the 16 MBTI categories does he/she falls into.

It is worth pointing out that MBTI is not designed to serve a clinical purpose. It is influenced by both nature and nurture, and may change throughout an individual's life course. A person's scores on the MBTI should be used as means to guide us making better decisions, and should not be used as labels.

Reference

Alexander, B. (2021, November 17). *Does Bert Need Clean Data? part 2 - classification.* Alexander Bricken. Retrieved April 22, 2022, from https://bricken.co/nlp_disaster_tweets_2/

Shperber, G. (2019, November 5). *A gentle introduction to doc2vec*. Medium. Retrieved April 22, 2022, from https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e

Stein, R., & Swan, A. B. (2019). Evaluating the validity of Myers‐Briggs Type Indicator theory: A teaching tool and window into intuitive psychology. *Social and Personality Psychology Compass*, *13*(2), e12434.

Sambursky, V. (2022, February 9). *Myers-Briggs test:* Limitations *and need for a better diagnostic tool* . Endominance. Retrieved April 22, 2022, from https://www.endominance.com/myers-briggs-test-limitations-and-need-for-a-better-diagnostic-tool/

Riggio , R. E. (2014, February 21). *The truth about Myers-Briggs types*. Psychology Today. Retrieved April 22, 2022, from https://www.psychologytoday.com/us/blog/cutting-edge-leadership/201402/the-truth-about-myers-briggs-types

Pittenger, D. J. (2005). Cautionary comments regarding the Myers-Briggs type indicator. *Consulting Psychology Journal: Practice and Research*, *57*(3), 210.

Boyle, G. J. (1995). Myers‐Briggs type indicator (MBTI): some psychometric limitations. *Australian Psychologist*, *30*(1), 71-74.

Geyer, P. (2009). Understanding the MBTI® and personality type. *Retrieved February*, *12*, 2010.

Lindecrantz, E., Gi, M. T. P., & Zerbi, S. (2020). Personalizing the Customer Experience: Driving Differentiation in Retail. *McKinsey Insights (March). Erişim Adresi https://www. mckinsey. com/industries/retail/our-insights/personalizing-the-customerexperience-driving-differentiation-in-retail*.

Evans, M. (2021, May 18). *The 4 'A's of marketing to different personality types*. Setup®. Retrieved April 24, 2022, from https://setup.us/blog/4-as-to-marketing-to-personality-types

Hunt, V., Layton, D., & Prince, S. (2021, March 12). *Why diversity matters*. McKinsey & Company. Retrieved April 24, 2022, from https://www.mckinsey.com/business-functions/people-and-organizational-performance/our-insights/why-diversity-matters