****

**Assessment Report**

on

**Personality Prediction using MBTI Dataset**

submitted as partial fulfillment for the award of

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AI)**

By

Group No. 7

Anwesha Singhal (202401100300061)

Animesh Pratap Singh (20240100300043)

Adarsh Kumar(202401100300012)

Akul Sharma (202401100300027)

Ashish Gupta (20240110030074)

Section: A

**Under the supervision of**

Mr. Bikki Gupta

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

1. **Introduction**

This project aims to predict personality types based on users’ written text using Natural Language Processing (NLP). Using the MBTI (Myers-Briggs Type Indicator) dataset, which includes 16 personality types, we apply text preprocessing, feature extraction, and machine learning models to classify user types. The goal is to explore how effectively language use reflects personality traits and demonstrate the potential of NLP in psychological profiling.

1. **Problem Statement**

Project Question: Use NLP techniques to classify personality types based on users? written text.

Implement text preprocessing and natural language understanding.

Dataset Link: https://www.kaggle.com/datasets/datasnaek/mbti-type

**3. Objectives**

· **To preprocess and clean user-generated text** for effective analysis using NLP techniques.

· **To extract meaningful features** from text using methods like TF-IDF or embeddings.

· **To classify personality types** using machine learning models based on the MBTI framework.

· **To evaluate model performance** using metrics such as accuracy and F1-score.

· **To explore the relationship between language patterns and personality traits.**

**4. Methodology**

### 1 Data Collection

The MBTI 5000 dataset was sourced from Kaggle. It includes user-generated text posts along with their corresponding MBTI personality types.

### 2 Text Preprocessing

All text data was cleaned using standard NLP techniques. This included converting to lowercase, removing punctuation, URLs, and stopwords, as well as applying tokenization and lemmatization to normalize the text for analysis.

### 3 Label Encoding

Each MBTI type was split into four binary categories:

· Introversion (I) / Extraversion (E)

· Intuition (N) / Sensing (S)

· Thinking (T) / Feeling (F)

· Judging (J) / Perceiving (P)

This allowed the model to approach the problem as four separate binary classification tasks.

### 4 Feature Extraction

Text data was transformed into numerical vectors using **TF-IDF** (Term Frequency-Inverse Document Frequency), which captures the importance of words in each user's posts.

### 5 Model Training

Machine learning algorithms such as **Logistic Regression** and **Support Vector Machines (SVM)** were used. Separate models were trained for each of the four MBTI dimensions. The dataset was split into training and test sets for evaluation.

### 6 Evaluation

Performance of each model was measured using classification metrics including **accuracy**, **precision**, **recall**, and **F1-score**. Confusion matrices were also analyzed to better understand the prediction results.

**5. Data Preprocessing**

The MBTI dataset text was cleaned by converting to lowercase, removing URLs, punctuation, and stopwords. Tokenization and lemmatization were applied to standardize words.

MBTI types were split into four binary labels (I/E, N/S, T/F, J/P) to simplify classification. The cleaned data was then split into training and test sets to evaluate model performance.

**6. Model Implementation**

Machine learning models were trained to classify each of the four MBTI personality dimensions separately. Text features were extracted using TF-IDF vectorization. Models including Logistic Regression and Support Vector Machines (SVM) were implemented due to their effectiveness in text classification tasks.

The dataset was split into training and testing sets, and each model was trained on the training data. Hyperparameters were tuned using cross-validation to optimize performance. Finally, the models were evaluated on the test set using accuracy and F1-score metrics.

**7. Evaluation Metrics**

To assess the performance of the personality prediction models, the following metrics were used:

**Accuracy:** Measures the overall percentage of correctly classified samples.

**Precision:** Indicates the proportion of true positive predictions among all positive predictions.

**Recall:** Measures the ability of the model to identify all relevant instances (true positives).

**F1-Score:** The harmonic mean of precision and recall, providing a balance between them.

**Confusion Matrix:** Shows detailed counts of true positives, true negatives, false positives, and false negatives for deeper error analysis.

**8. Results and Analysis**

The trained models showed varied performance across the four MBTI dimensions. Generally, the classifiers achieved moderate accuracy, with better results on distinguishing **Introversion/Extraversion (I/E)** and **Thinking/Feeling (T/F)** traits compared to the others.

**Accuracy** scores ranged from approximately 65% to 75% depending on the personality dimension.

**F1-scores** indicated a balanced precision and recall, though some classes exhibited slight imbalance in predictions.

The **confusion matrices** revealed that the model occasionally confused similar personality traits, suggesting overlapping language patterns between certain MBTI categories.

**9. Conclusion**

This project demonstrated the feasibility of predicting MBTI personality types from users’ written text using NLP and machine learning techniques. Through text preprocessing, feature extraction, and classification, the models achieved reasonable accuracy, particularly in distinguishing certain personality dimensions like Introversion/Extraversion and Thinking/Feeling.

However, some personality traits were more challenging to classify, indicating that subtle differences in language use may require more advanced models or richer text representations. Future work could explore deep learning approaches such as BERT embeddings to improve prediction performance.

Overall, this study highlights the potential of combining psychological theory with NLP for personality profiling, which can be valuable in applications like personalized recommendations and social analysis.

1. **References**

· **MBTI Dataset**  
MBTI 5000 dataset. Kaggle. Available at: <https://www.kaggle.com/datasnaek/mbti-type>

· **pandas** (import pandas as pd)

For loading and handling the dataset as a DataFrame

· **re**

For regular expressions used in text cleaning (removing URLs, special characters).

· **nltk** (from nltk.corpus import stopwords, from nltk.stem import WordNetLemmatizer)

For natural language processing tasks such as removing stopwords and lemmatization (normalizing words)

