MBTI Classification Model Based on NLP

Tsai Hsin Yu Tian Yuhang Liu Kexuan Wang Yixuan Huang Xinrui

Abstract

This project predicts MBTI types from writtentext. Using data from Kaggle and Alibaba Cloud'sTianchi platform, we explored several models in-cluding Logistic Regression, BERT, RoBERTaand T5, testing both direct 16-class classificationand multi-binary classification. Our results re-vealed that BERT with 16-class classification per-formed best. RoBERTa showed comparable re-sults but was more sensitive to data quality andhyperparameters, while T5's generative approachwas less effective. The BERT 16-class model wasthen deployed via an APl through Hugging Face,enabling real-time MBTI predictions based onuser text.

1. Introduction

In contemporary society, the need for self-understanding and personality insights has grown. Individuals seek to recognize their behavioral patterns, emotional responses, and interaction styles to make better decisions in areas like career, relationships, and personal growth. The Myers-Briggs Type Indicator (MBTI) is one of the most widely used tools for personality assessment, categorizing individuals into 16 personality types based on four dimensions: Extraversion (E) vs. Introversion (I), Intuition (N) vs. Sensing (S), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P). Each type, such as INFP or ESTJ, reflects the cognitive and decision-making preferences of an individual.

However, traditional MBTI assessments rely on self-reported questionnaires, which can be biased by the respondent's emotional state or self-perception, limiting their real-world accuracy. To address this, our project proposes a language-based personality prediction system that provides dynamic, context-aware insights. We start with a baseline Logistic Regression model and then fine-tune a pre-trained BERT model for improved performance. We also explore other models to enhance results. In addition to predicting the overall 16 MBTI types, we treat the four personality dimensions (E/I, N/S, T/F, J/P) as separate binary classification tasks, training four additional models for comparative analysis. Finally, we deploy an API for the best-performing 16-type classification model, allowing users to input text and receive real-time MBTI predictions. This approach leverages the real-time analytical power of language models to offer a more dynamic and practical alternative to traditional MBTI assessments.

2. Business Value

This project creates substantial business value through data-driven decision-making in three core areas: recruitment, customer engagement, and team optimization. For recruitment, HR professionals can leverage the MBTI prediction API to assess candidates' personality traits by analyzing their written responses like self-introductions or answers to open-ended questions. The API's language pattern analysis offers valuable insights into work styles and cultural fit, resulting in better hiring decisions, improved role alignment, and lower turnover. In customer operations, integrating the API enables more personalized interactions. By analyzing customer inputs from feedback, chats, or surveys, businesses can identify personality preferences and adjust their communication approach accordingly—whether factual or emotional—to boost satisfaction, loyalty, and conversions. For team management, the API evaluates personality types through members' written input, including self-assessments and collaboration experiences. These insights help managers build balanced teams, combine complementary strengths, and reduce conflicts, creating more productive work environments.

Across these applications, personality analytics delivers measurable improvements in hiring quality, customer relations, and team performance, driving long-term organizational success through intelligent, data-supported strategies.

3. Data Collection and Preprocessing

3.1. Data Description

We leverage two complementary data sources to build a robust MBTI classification model. Our primary dataset is the Kaggle "MBTI Personality Type Twitter Dataset" comprising 7,800 user samples; each record contains an MBTI label (four-letter code) and the user's historical tweets delimited by "|||". To mitigate overfitting due to limited Twitter data, we incorporate a secondary corpus from Alibaba Cloud's Tianchi platform, drawn from the Personality-

Cafe forum and containing over 8,600 similarly structured samples.

3.2. Data Preprocessing & EDA

Data preprocessing began with rigorous cleaning and normalization to ensure consistency across both sources. We first removed all records containing null values and standardized MBTI labels (lowercased, four-letter codes) so that Kaggle and Tianchi data adhere to an identical format. To facilitate embedding and mitigate bias toward users with extensive posting histories, we exploded each multi-post record into individual entries, each comprising a single post and its associated MBTI label.

All text was then lowercased and cleaned via regular expressions: HTML tags, URLs, user mentions, hashtags, emojis, and extraneous whitespace were stripped, while alphanumeric characters and essential punctuation were retained. After merging the two datasets, exploratory visualizations (label distribution bar charts and post-length histograms) revealed that the top six MBTI types (INFP, INFJ, INTP, INTJ, ENFP, ENTP) constitute over 50% of the data and that similar imbalances persist at the single-letter dimension. This kind of data imbalance persist when separate each component letter of MBTI type. To address class imbalance and limit computational overhead, we under-sampled majority classes—randomly selecting an equal number of posts per MBTI type using a fixed random seed for reproducibility. Finally, we employed Hugging Face's 'bert-base-uncased' tokenizer to tokenize our posts, and generated attention masks and token type IDs. Additionally, a stratified 70/15/15 train/validation/test split is conducted to ensure unbiased model evaluation.

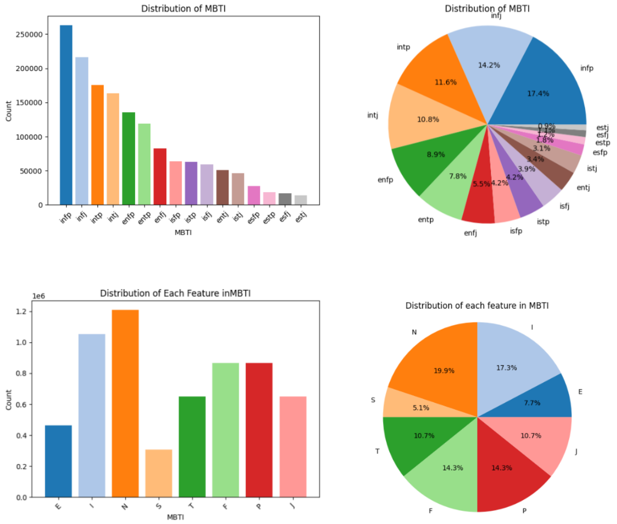


Figure 1. Distribution of MBTI and Features

4. Model Development and Evaluation

4.1. Methodology

At the initial planning stage of this project, we designed MBTI personality prediction as a standard 16-class classification task, where each MBTI type (e.g., "INFP", "ESTJ") was treated as a distinct categorical label. This straightforward formulation allowed for the direct application of conventional multi-class classification models. However, from a theoretical perspective, this approach inherently treats all personality types as mutually exclusive and equidistant, ignoring the underlying structured composition of MBTI types along four orthogonal psychological dimensions: Energy (E/I), Information (N/S), Decision (T/F), and Lifestyle (P/J).

Recognizing these structural characteristics, we anticipated potential limitations in the 16-class setup. For instance, while "ENTP" and "ENFP" differ solely in the T/F dimension, the 16-class classifier lacks any mechanism to explicitly model their proximity. Additionally, semantically invalid combinations like "EINF" could not be ruled out in this flat classification approach. These considerations suggested that the model might suffer from impaired interpretability and an inability to leverage MBTI's inherent compositional structure.

To address these issues, we planned a reformulation of the task into a multi-binary classification framework. Instead of predicting a single 16-type label, we devised four independent binary classifiers, each targeting one MBTI dimension. In this setup, each classifier specializes in a single psychological dichotomy—Extraversion vs. Introversion, Intuition vs. Sensing, Thinking vs. Feeling, and Perceiving vs. Judging—thereby enabling dimension-specific learning and interpretability. This design, conceptually akin to a Mixture of Experts, was aimed at enhancing model transparency and providing fine-grained diagnostic capabilities.

The training labels for each binary classifier were generated by decomposing the original 16-type labels into a set of four binary indicators. For instance, the type "ENFP" was mapped to the vector [1, 1, 0, 1], representing the presence of traits E, N, F, and P. This decomposition was intended to allow the models to focus on identifying linguistic patterns associated with each psychological dimension independently, potentially improving both model interpretability and robustness.

At inference time, the four classifiers independently produced probability scores for their respective dimensions. Thresholding these scores (typically at 0.5) enabled binary decisions, which were then recombined to generate the predicted MBTI type (e.g., [1, 1, 0, 1] → "ENFP"). This modular strategy also allowed us to examine whether the model's outputs adhered to MBTI structural constraints (e.g., preventing invalid types like “EINF”).

However, we recognized that this modeling choice introduced a trade-off: while the multi-binary framework promotes modular analysis, it assumes independence between dimensions. In reality, MBTI types exhibit inter-dimensional dependencies and non-uniform co-occurrence patterns. Some combinations, such as “INFP” or “ENFJ,” are more frequent and coherent than others. Without explicitly modeling these joint distributions, the binary classifiers might produce plausible but statistically improbable types.

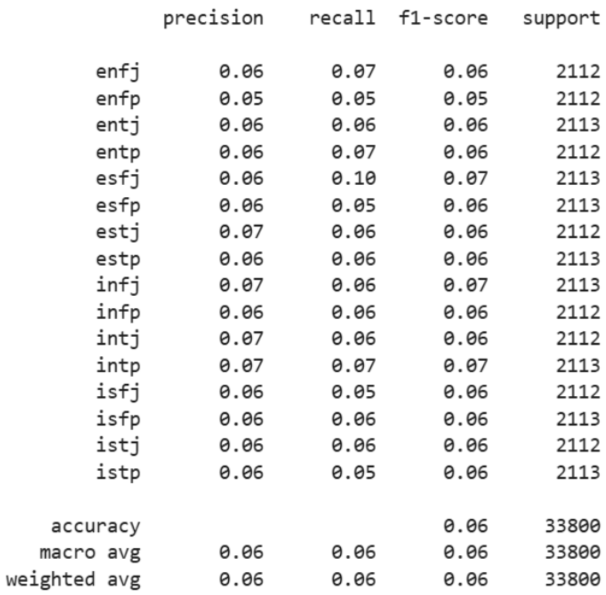
In parallel, we continued to explore improvements by introducing stronger pretrained models such as RoBERTa (Robustly Optimized BERT Pretraining Approach). Building on the limitations observed in BERT, RoBERTa offered a more powerful language encoder with enhanced pretraining techniques. The idea was to examine whether end-to-end fine-tuning of a larger model could overcome some of the weaknesses identified in our earlier configurations. Following this, we also explored T5 (Text-to-Text Transfer Transformer), leveraging its flexible text-to-text paradigm to reframe personality prediction as a sequence generation task, although practical constraints hindered the realization of its full potential within our setting.

Throughout this methodological progression, we consistently balanced between maximizing model expressiveness and maintaining interpretability, planning each subsequent step based on both theoretical motivations and empirical findings from previous phases.

4.2. Logistic Regression (Baseline)

We chose Logistic Regression as our baseline model for predicting MBTI personality types for three main reasons. First, its straightforward design makes it easy to implement and interpret, while its clear coefficients help reveal the influence of key features in the text data. Second, compared to more complex models, Logistic Regression has relatively low computational demands, allowing us to quickly establish a performance benchmark before exploring more advanced alternatives. Finally, if its effectiveness in binary and multi-class classification tasks is proven, it will be a reliable foundation for modeling categorical data such as MBTI types, especially when dealing with text-based inputs.

First, let’s examine the performance of the 16-class MBTI type prediction model. As shown in the results, the model achieves an extremely low accuracy of just 6.24% - even worse than random guessing (which would yield 6.25% accuracy for 16 balanced classes). The model also fails uniformly across all types, showing nearly identical precision, recall, and F1 scores ranging between 0.05 and 0.07, with no type demonstrating even marginally better performance than others. While the ’ESFJ’ type shows a slightly higher recall of 0.10, this minor variation appears to be random rather than representing meaningful differentiation.



```
              precision    recall  f1-score   support

        enfj       0.06      0.07      0.06      2112
        enfp       0.05      0.05      0.05      2112
        entj       0.06      0.06      0.06      2113
        entp       0.06      0.07      0.06      2112
        esfj       0.06      0.10      0.07      2113
        esfp       0.06      0.05      0.06      2113
        estj       0.07      0.06      0.06      2112
        estp       0.06      0.06      0.06      2113
        infj       0.07      0.06      0.07      2113
        infp       0.06      0.06      0.06      2112
        intj       0.07      0.06      0.06      2112
        intp       0.07      0.07      0.07      2113
        isfj       0.06      0.05      0.06      2112
        isfp       0.06      0.06      0.06      2113
        istj       0.06      0.06      0.06      2112
        istp       0.06      0.05      0.06      2113

    accuracy                           0.06     33800
   macro avg       0.06      0.06      0.06     33800
weighted avg       0.06      0.06      0.06     33800
```

Figure 2. Logistic Regression Classification Report (16-Class)

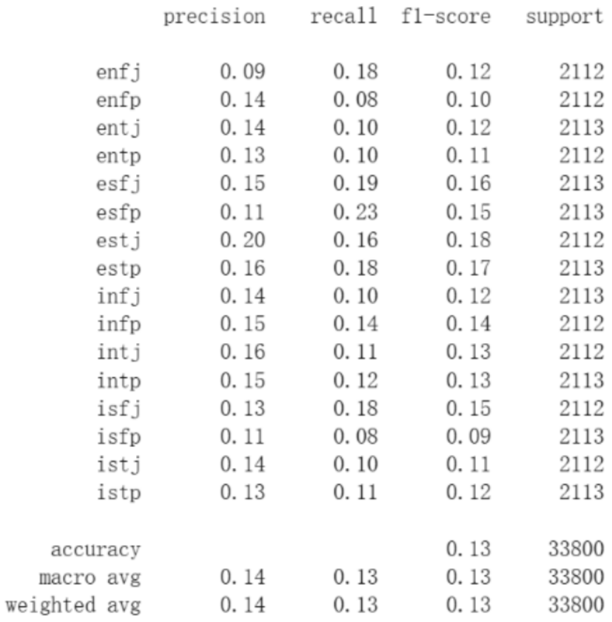
Next, let’s examine the performance of the classifiers specific to each dimension. From the comparison table of classification reports below, we can observe that all classifiers achieved accuracy scores hovering around 0.5, indicating that the model essentially lacks discriminative power and performs no better than random guessing - equivalent to a coin flip. Furthermore, the comparable performance between Class 0 and Class 1 suggests the model shows no significant preference or bias toward either class in its predictions.

Dimension	Accuracy	F1 (Class 0)	F1 (Class 1)
E/I	49.8%	0.48	0.51
N/S	49.8%	0.51	0.49
T/F	50.2%	0.49	0.51
P/J	49.5%	0.48	0.51

Table 1. Logistic Regression Classification Report (Binary)

4.3. BERT Based 16 Classification Model

Realizing logistic regression does not have strong predictive ability, we then further explored Bidirectional Encoder Representation (BERT) model to conduct 16 categories classification task. In the model building stage, we tried different hyperparameters for the model. Worring overfitting and limited computational resources, we applied 2 layers neural network with a dropout rate of 0.1 and the AdamW optimizer with learning rate of 3e-5 and weight decay of 1e-6. However, we end up finding our model tend to underfitting when evaluating on the validation set. Therefore, we finally employed a 3 layer neural network with dropout rate of 0.05 with no freezing layers. To accelerate the convergence speed, we adjusted the hyperparameters of AdamW optimizer to a learning rate of 2e-5 and a weight decay of 1e-3.



```
              precision    recall  f1-score   support

        enfj       0.09      0.18      0.12      2112
        enfp       0.14      0.08      0.10      2112
        entj       0.14      0.10      0.12      2113
        entp       0.13      0.10      0.11      2112
        esfj       0.15      0.19      0.16      2113
        esfp       0.11      0.23      0.15      2113
        estj       0.20      0.16      0.18      2112
        estp       0.16      0.18      0.17      2113
        infj       0.14      0.10      0.12      2113
        infp       0.15      0.14      0.14      2112
        intj       0.16      0.11      0.13      2112
        intp       0.15      0.12      0.13      2113
        isfj       0.13      0.18      0.15      2112
        isfp       0.11      0.08      0.09      2113
        istj       0.14      0.10      0.11      2112
        istp       0.13      0.11      0.12      2113

    accuracy                           0.13     33800
   macro avg       0.14      0.13      0.13     33800
weighted avg       0.14      0.13      0.13     33800
```

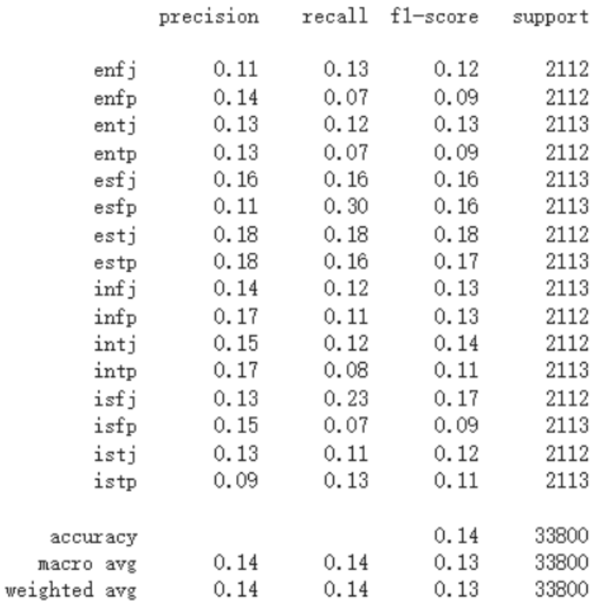
Figure 3. BERT Classification Report on Full Dataset (16-Class)

The performance of BERT 16 classification model is slightly, but limited, better than logistic regression. Our model has a slightly large cross entropy loss of 2.78, which indicates our model fails to learn significant patterns to make classification during the training process. The F-1 score of 0.1325, although two-fold compare with the probability of random guess, still not accurate enough, confirming the limited classification ability. Additionally, a F-1 score of 0.1325, a precision score of 0.1398, and a recall score of 0.1346 indicates a high false positive and false negative classifications. Dive into the detail, we find the model has slightly strong ability in predicting 'ESTJ' MBTI type with an precison of 0.2 and a recall of 0.16 meanwhile performs the worst when predicting 'ENFJ' MBTI type. This might indict the quality of two types of MBTI dataset could be greatly different. Besides these two MBTI types, all other types have accuracy rate, Recall, and F-1 score between 0.1 and 0.2. Therefore, although the BERT 16 classification model is slightly better than random guess and logistic regression model, the current performance indicates this model is still far from realistic application usage.

At this stage, we hypothesized that the model's suboptimal performance might stem from a probability distributional mismatch between the two merged datasets. Specifically, differences in data collection timeframes, geographic origins, and potential sampling biases could have led to significant divergences in word usage probabilities between the two datasets, even among users sharing the same MBTI personality type. In the worst-case scenario, the feature distributions for the same MBTI in one dimension could even exhibit reversed patterns across datasets. Simply concatenating these two datasets without accounting for such differences might hinder the model's ability to learn any coherent underlying patterns from either source.

To test this hypothesis, we conducted an additional experiment where we trained and evaluated the same model on a single dataset. The result showed that the model's performance on the single dataset was not significantly better, with slightly better accuracy but worse F1 Score, than on the merged dataset, suggesting that our initial concern was unnecessary. The relatively poor model performance did not originate from distributional shifts between the two datasets. In fact, merging the datasets may be beneficial, likely due to the scale effect and the increased diversity of training samples, which can help regularize the model and mitigate overfitting.

Therefore, for all subsequent experiments, we continue to use the merged dataset.



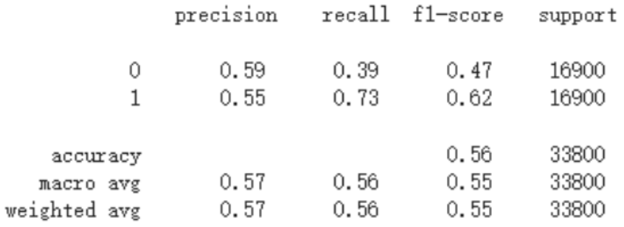
```
              precision    recall  f1-score   support

        enfj       0.11      0.13      0.12      2112
        enfp       0.14      0.07      0.09      2112
        entj       0.13      0.12      0.13      2113
        entp       0.13      0.07      0.09      2112
        esfj       0.16      0.16      0.16      2113
        esfp       0.11      0.30      0.16      2113
        estj       0.18      0.18      0.18      2112
        estp       0.18      0.16      0.17      2113
        infj       0.14      0.12      0.13      2113
        infp       0.17      0.11      0.13      2112
        intj       0.15      0.12      0.14      2112
        intp       0.17      0.08      0.11      2113
        isfj       0.13      0.23      0.17      2112
        isfp       0.15      0.07      0.09      2113
        istj       0.13      0.11      0.12      2112
        istp       0.09      0.13      0.11      2113

    accuracy                           0.14     33800
   macro avg       0.14      0.14      0.13     33800
weighted avg       0.14      0.14      0.13     33800
```

Figure 4. BERT Classification Report on Single Dataset (16-Class)

4.4. BERT Based Multi-Binary Classification Model

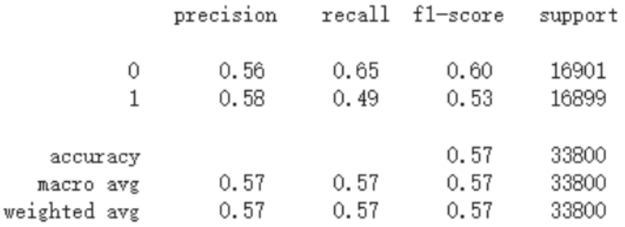
To ensure a fair comparison, each binary classifier was trained with the same data, hyperparameters, and evaluation settings as the 16-class model, including identical BERT encoder configurations, optimizers, batch size, and epochs. Model performance was evaluated on the same test set. Although each binary model shares the same number of parameters as the 16-class classifier in BERT layers, we expected improved performance with this setup.

Figures 5–8 show the performance of each dimension-specific classifier. The Energy dimension had high recall for Extraverts (0.7314) but lower precision (0.5454), indicating frequent misclassification of Introverts as Extraverts, likely due to more obvious Extraversion features in user text. The Information dimension showed the opposite: higher precision but lower recall for Intuitive types, suggesting the model was more conservative but accurate when predicting Intuition.



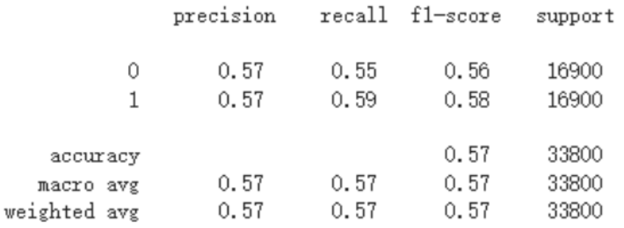
```
              precision    recall  f1-score   support

           0       0.59      0.39      0.47     16900
           1       0.55      0.73      0.62     16900

    accuracy                           0.56     33800
   macro avg       0.57      0.56      0.55     33800
weighted avg       0.57      0.56      0.55     33800
```

Figure 5. BERT Classification Report (E/I)



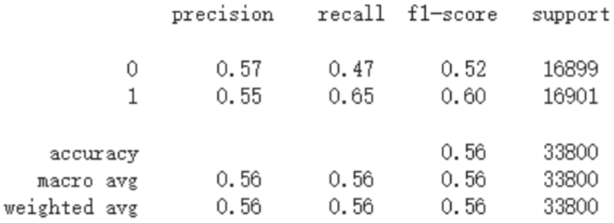
```
              precision    recall  f1-score   support

           0       0.56      0.65      0.60     16901
           1       0.58      0.49      0.53     16899

    accuracy                           0.57     33800
   macro avg       0.57      0.57      0.57     33800
weighted avg       0.57      0.57      0.57     33800
```

Figure 6. BERT Classification Report (N/S)



```
              precision    recall  f1-score   support

           0       0.57      0.55      0.56     16900
           1       0.57      0.59      0.58     16900

    accuracy                           0.57     33800
   macro avg       0.57      0.57      0.57     33800
weighted avg       0.57      0.57      0.57     33800
```

Figure 7. BERT Classification Report (T/F)



```
              precision    recall  f1-score   support

           0       0.57      0.47      0.52     16899
           1       0.55      0.65      0.60     16901

    accuracy                           0.56     33800
   macro avg       0.56      0.56      0.56     33800
weighted avg       0.56      0.56      0.56     33800
```

Figure 8. BERT Classification Report (J/P)

The Decision dimension yielded the most balanced results, with both precision and recall near 0.57–0.59, suggesting that the contrast between logical reasoning and emotional expression is more linguistically distinct and evenly distributed in the dataset. The Lifestyle dimension mirrored the Energy dimension's asymmetry, with higher recall but lower precision for the perceiving class.

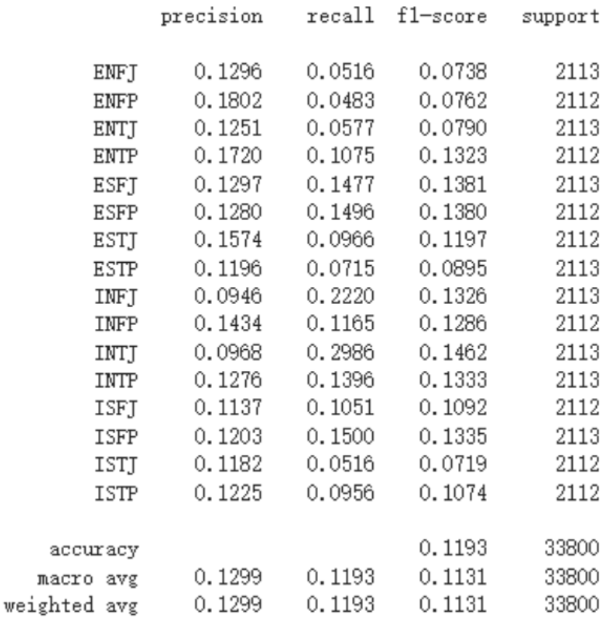
We further evaluated the joint performance of the integrated model by recombining the four binary predictions into complete MBTI types and comparing them with the ground truth labels. The combined model achieved an overall accuracy of 11.93%, which is consistent with the expected compound error rate derived from the individual dimension accuracies (approximately 10.2% from the product of four independent accuracies), although the integrated model actually has nearly three times more parameters than the 16-class classifier. This result confirms that errors made at the dimension level propagate and amplify when predicting full personality types.

We further evaluated the joint performance of the integrated model by recombining the four binary predictions into complete MBTI types and comparing them with the ground truth labels. The combined model achieved an overall accuracy of 11.93%, which is consistent with the expected compound error rate derived from the individual dimension accuracies (approximately 10.2% from the product of four independent accuracies), although the integrated model actually has nearly three times more parameters than the 16-class classifier. This result confirms that errors made at the dimension level propagate and amplify when predicting full personality types.

Detailed analysis of type-level performance reveals additional insights. Certain types, such as INFJ and INTJ, exhibited notably higher recall (0.222 and 0.299, respectively), indicating a tendency of the model to overpredict these combinations. However, their low precision values (0.0946 and 0.0968) suggest a high rate of false positives—an issue not observed in the 16-class model—highlighting the risks of compounding misclassifications in the binary setup.

On the other hand, types such as ENFP and ISTJ showed particularly low F1 scores, likely due to the instability of individual dimension predictions. Interestingly, types like ISFP and ESFJ achieved more balanced performance, possibly because the underlying dimensions for these types are more consistently expressed in linguistic patterns.

Overall, while the multi-binary decomposition aligns better with the theoretical underpinnings of MBTI and supports targeted diagnostics, it does not outperform the 16-class classifier in terms of full-type accuracy. Moreover, its assumption of inter-dimensional independence leads to compounded prediction errors. Future work may benefit from incorporating joint modeling techniques—such as structured prediction, conditional dependency modeling, or probabilistic coherence constraints—to preserve both modular interpretability and inter-dimensional consistency.



```
              precision    recall  f1-score   support

        ENFJ     0.1296    0.0516    0.0738      2113
        ENFP     0.1802    0.0483    0.0762      2112
        ENTJ     0.1251    0.0577    0.0790      2113
        ENTP     0.1720    0.1075    0.1323      2112
        ESFJ     0.1297    0.1477    0.1381      2113
        ESFP     0.1280    0.1496    0.1380      2112
        ESTJ     0.1574    0.0966    0.1197      2112
        ESTP     0.1196    0.0715    0.0895      2113
        INFJ     0.0946    0.2220    0.1326      2113
        INFP     0.1434    0.1165    0.1286      2112
        INTJ     0.0968    0.2986    0.1462      2113
        INTP     0.1276    0.1396    0.1333      2113
        ISFJ     0.1137    0.1051    0.1092      2112
        ISFP     0.1203    0.1500    0.1335      2113
        ISTJ     0.1182    0.0516    0.0719      2112
        ISTP     0.1225    0.0956    0.1074      2112

    accuracy                         0.1193     33800
   macro avg     0.1299    0.1193    0.1131     33800
weighted avg     0.1299    0.1193    0.1131     33800
```

Figure 9. BERT Classification Report (Integrated Model)

4.5. RoBERTa Expolration

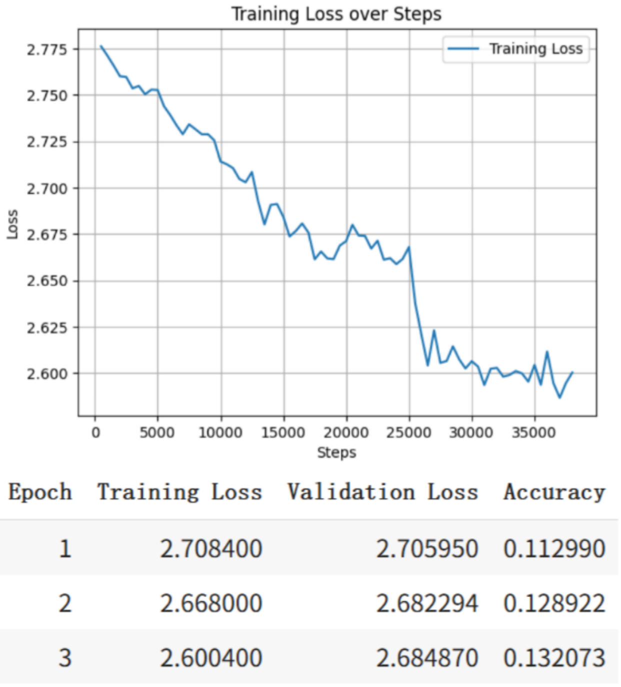
To explore more alternative model selections, we tried the RoBERTa model (Robustly Optimized BERT Pre-training Approach) to perform a 16-class MBTI personality classification based on user text posts. RoBERTa is an advanced transformer-based language model that improves upon BERT by removing the next-sentence prediction objective and training with larger mini-batches and longer sequences. It has shown strong performance in various downstream tasks due to its pre-training on a large corpus with dynamic masking strategies.

The model was fine-tuned on our resampled dataset, using a multiclass classification head (number of labels=16) on top of the roberta-base encoder. The learning rate was set to 2e-5 with linear scheduling and a warm-up ratio of 0.1, and the training was carried out over 3 epochs. Accuracy was used as the primary evaluation metric.

During training, we observed a gradual decrease in training loss from 2.7084 in epoch 1 to 2.6004 by epoch 3. The validation loss remained relatively stable, fluctuating slightly between 2.7059 and 2.6849. The validation accuracy showed a modest but consistent upward trend, improving from 11.30% in the first epoch to a final value of 13.21% by the third epoch.

The training loss curve further confirms convergence, with a noticeable drop around the midpoint of training and gradual stabilization thereafter. This suggests that while the model is being studied, it might be approaching a performance ceiling under the current configuration.

Despite using a powerful pretrained model, the resulting accuracy did not exceed 13.3%. This can be attributed to several factors such as label complexity (16 MBTI classes), the relatively limited input length, and subtle linguistic features that may not be fully captured by token-level representations.



Epoch	Training Loss	Validation Loss	Accuracy
1	2.708400	2.705950	0.112990
2	2.668000	2.682294	0.128922
3	2.600400	2.684870	0.132073

Figure 10. RoBERTa Performance

In comparison, the BERT-based embedding combined with an MLP classifier achieved slightly better results in terms of accuracy. This discrepancy may be due to differences in modeling approaches. In the BERT + MLP setup, BERT serves as a static feature extractor, providing relatively stable semantic representations, just like the rich and diverse daily words usage of online posts, which is more suitable for tasks with smaller label spaces. In contrast, RoBERTa is fine-tuned end-to-end, with larger model capacity and stronger expressive power, but also more prone to overfitting or unstable training when hyperparameters like learning rate or training epochs are not well-optimized.

Furthermore, the BERT + MLP framework is structurally lighter and easier to control during training, leading to more stable overall performance. However, RoBERTa training is more sensitive to data quality and hyperparameter settings, which may have limited its performance in this particular task.

4.6. T5 Exploration

Following our previous experiments with BERT and RoBERTa models, we further explored the use of the T5 model for this task. Unlike encoder-only models like BERT and RoBERTa, T5 frames the entire problem in a pure text-to-text format, which offers greater flexibility in designing prompts and aligning inputs and outputs seamlessly. Its generative nature also allows for more natural handling of structured label outputs, making it a promising alternative for this kind of classification task.

In this phase, we adopted a similar task setup: the dataset was stratified and split into training, validation, and test sets, and each post was paired with either a full MBTI type label or a label corresponding to one of the four dimensions. Using carefully designed prompts, we trained the T5 model to predict the full personality type or individual dimensions. The model, tokenizer, and configuration were based on the pre-trained ”t5-base,” with minor modifications such as ”dropout=0.1” and ”batch-size=16”. Tokenization and preprocessing were performed accordingly, and training was carried out using Hugging Face’s Trainer. This continuation of work builds directly upon our earlier approaches, providing a comparative perspective between discriminative and generative modeling strategies for MBTI prediction.

Despite careful prompt design and data preparation, the T5 model’s predictions did not meet expectations. Although the model showed a low loss at the end of training, its predictions were still inaccurate. This discrepancy may be due to the focus of the model on generating fluent text rather than precise classifications. Additionally, T5 uses cross-entropy loss, which works well for generative tasks but may not align perfectly with classification tasks, affecting its ability to predict discrete MBTI types.

While T5’s architecture is powerful, its performance in MBTI prediction was limited by memory and resource constraints during training, which hindered its ability to capture the complex relationships between posts and MBTI labels. Furthermore, the inherent complexity of predicting MBTI types involves understanding nuanced psychological traits, making the task more challenging. To improve accuracy, future experiments could explore domain-specific models, hybrid approaches, or preprocessing techniques like sentiment analysis. Further adjustments to the hyperparameters and the expansion of the training dataset could also enhance prediction performance.

In conclusion, while the T5-base model offers flexibility in generating text and handling structured outputs, its performance in MBTI prediction was less accurate compared to our previous experiments with BERT and RoBERTa. Despite achieving a low loss during training, T5’s generative nature, focused on producing fluent text rather than precise classifications, led to inaccurate predictions. Unlike discriminative models like BERT and RoBERTa, T5 struggled to capture the complex relationships between posts and MBTI labels. Furthermore, computational limitations may have hindered its performance. This comparison suggests that, while generative models such as T5 offer flexibility, they may require more domain-specific adjustments for classification tasks. Future improvements might involve combining discriminative and generative models, refining preprocessing methods, and expanding the training set to better capture the nuances of MBTI prediction.

4.7. Result Analysis and Model Selection

To comprehensively evaluate alternative modeling strategies for 16-class MBTI personality classification, we compared the performance of baseline and several transformer-based architectures, including a BERT-based multi-binary classifier, a RoBERTa fine-tuned model, and a T5-based model, against the final selected BERT 16-class classifier. Although the performance of none of them was satisfactory.

Model	Number of parameters	F1 on test set	Accuracy on test set
Logistic Regression	12,304	0.0621	0.0624
BERT 16-class classifier	110M	**0.1325**	**0.1346**
BERT multi-binary model	4*110M	0.1131	0.1193
RoBERTa	125M	0.1286	0.1330
T5	220M	Na	Na

Table 2. Models’ Performance Comparison

The BERT multi-binary classifier, which treats each MBTI dimension as an independent binary task, showed lower overall performance compared to the direct 16-class classification approach. This is likely due to the loss of inter-dimensional dependencies between MBTI traits, as treating each axis separately ignores the underlying correlations embedded in user text patterns. Additionally, the cumulative parameter size across four independent classifiers led to increased model complexity without proportional gains in generalization, making the training process less efficient and more susceptible to overfitting.

The RoBERTa model, despite its enhanced pretraining strategies and greater expressive capacity compared to BERT, achieved a comparable but slightly lower test accuracy relative to the BERT 16-class classifier. This may stem from RoBERTa's increased sensitivity to hyperparameters and data quality, especially in settings with a limited amount of nuanced linguistic input and a highly granular label space like MBTI. In contrast, the BERT-based model, with a slightly more controlled optimization process and smaller label smoothing effects, demonstrated more stable convergence behavior and better alignment with the task characteristics.

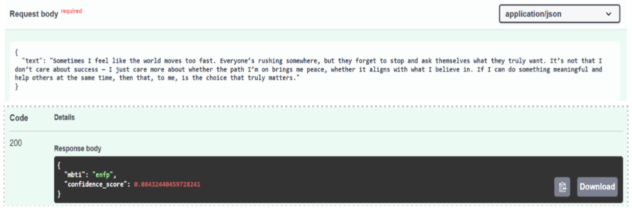
Explorations with the T5 model, which frames tasks in a text-to-text format, failed to meet performance expectations for this classification task. The inherent mismatch between T5's generative architecture and the discrete nature of classification objectives led to difficulties in optimization and effective learning. Unlike extractive or discriminative models, T5's formulation introduces additional variability and decoding complexities, which proved suboptimal under the current task requirements and dataset conditions.

In conclusion, by balancing performance stability, model complexity, and task alignment, the BERT 16-class classifier emerged as the most suitable choice for this MBTI classification task, achieving the best test set performance in both accuracy and F1 score.

5. API Deployment

After obtaining the best results from our top-performing BERT model, we proceeded to develop an MBTI classification API. To deploy the API, we created a Hugging Face Space, uploaded the necessary files, and accessed the API through a public link. This link leads to an automatically generated interactive documentation (Swagger UI), which allows users to test our API directly in the browser without writing any code. The primary objective of deploying this API is to enable external users to engage with and apply the outcomes of our project, thereby realizing the business value outlined in Section 3. The API is designed for ease of use—simply click this link, select "Try it out," and enter a sentence into the input box. For instance, in the example shown below, we input a sentence reflective of the INFP personality type. Upon clicking "Execute," the API returns the predicted MBTI type along with a confidence score. In this case, the model predicted ENFP with a confidence score of 0.08.

Although the confidence score is relatively low, the predicted result is still directionally consistent with the true type, suggesting that the model effectively captures key personality traits. This demonstrates the model's underlying potential, even under conditions of low certainty. Nonetheless, we are committed to continuously improving prediction accuracy and reliability. This initial deployment serves as a robust foundation for future development. Thanks to the API's modular design, we can upgrade the backend—such as incorporating larger training datasets or more sophisticated model architectures—without altering the interface. As a result, users can benefit from ongoing performance improvements without any disruption to their experience.



```
{
  "text": "Sometimes I feel like the world moves too fast. Everyone’s rushing somewhere, but they forget to stop and ask themselves what they truly want. It’s not that I don’t care about success — I just care more about whether the path I’m on brings me peace, whether it aligns with what I believe in. If I can do something meaningful and help others at the same time, then that, to me, is the choice that truly matters."
}
```

```
{
  "mbti": "enfp",
  "confidence_score": 0.08432440459728241
}
```

Figure 11. API deployment

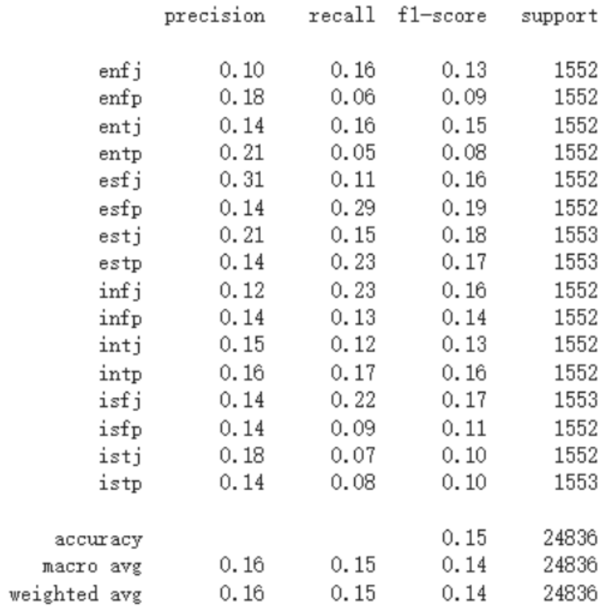
6. Limitation and Future Improvement

Due to the unsatisfactory models' performance, it is needed to discuss the limitations of this project. This may offer avenues for future research.

First, due to computational resource constraints, we addressed the issue of class imbalance through undersampling rather than employing more sophisticated techniques such as weighted loss functions. This decision inevitably resulted in the loss of a substantial amount of information and disproportionately amplified the influence of minority MBTI type publishers, potentially introducing bias into the model training process.

Second, regarding the input data, many online posts are relatively short in length, providing insufficient textual evidence to comprehensively represent a user's personality. Short posts limit the capacity of Transformer-based architectures to fully exploit their attention mechanisms, which are most effective when modeling long-range dependencies. Treating short posts as equally important observations alongside longer ones may have reduced the model's ability to capture deeper underlying personality patterns. For instance, in another training attempt of our 16-class classifier, we eliminated overly short posts, and the accuracy performance of the model improved slightly by 0.01.

Furthermore, each post was treated as an independent observation, which does not accurately reflect the real-world dynamics of user expression. In practice, individuals often convey their emotions, thoughts, and personalities across multiple posts within a temporal context, with potential emotional fluctuations and evolving viewpoints. Such higher-level contextual patterns are likely to be closely tied to stable personality traits. However, due to the lack of timestamp information associated with posts, we were unable to model these temporal dynamics or capture cross-post dependencies effectively.



```
              precision    recall  f1-score   support

        enfj       0.10      0.16      0.13      1552
        enfp       0.18      0.06      0.09      1552
        entj       0.14      0.16      0.15      1552
        entp       0.21      0.05      0.08      1552
        esfj       0.31      0.11      0.16      1552
        esfp       0.14      0.29      0.19      1552
        estj       0.21      0.15      0.18      1553
        estp       0.14      0.23      0.17      1553
        infj       0.12      0.23      0.16      1552
        infp       0.14      0.13      0.14      1552
        intj       0.15      0.12      0.13      1552
        intp       0.16      0.17      0.16      1552
        isfj       0.14      0.22      0.17      1553
        isfp       0.14      0.09      0.11      1552
        istj       0.18      0.07      0.10      1552
        istp       0.14      0.08      0.10      1553

    accuracy                           0.15     24836
   macro avg       0.16      0.15      0.14     24836
weighted avg       0.16      0.15      0.14     24836
```

Figure 12. BERT Classification Report (16-Class, after Filtering out too Short Posts)

Additionally, limitations in computational resources and time prevented us from thoroughly exploring the potential advantages of ensemble modeling. As discussed previously, the 16-class classifier and the multi-binary classifier demonstrated complementary strengths across different MBTI types. Integrating these models through ensemble methods could leverage their respective advantages and might lead to unexpectedly strong performance, which remains an open avenue for future exploration.

Moreover, hyperparameter optimization for the RoBERTa model was limited. A more extensive and systematic search over learning rates, batch sizes, warmup strategies, and regularization techniques could potentially unlock better performance, given RoBERTa's known sensitivity to fine-tuning configurations.

Another notable limitation concerns the modality of input data. In today's digital communication landscape, users increasingly express their emotions and thoughts not only through text but also through emojis, images, videos, and other multimodal content. By focusing exclusively on text-based inputs, our study constrained the model's ability to access the full richness of user expression, thereby setting an upper bound on achievable performance.

Finally, from an epistemological perspective, it is important to acknowledge that personality is often considered dynamic and context-dependent rather than fixed and immutable. Our study inherently assumes that posts at different time points are reflective of a user's static, self-reported MBTI type. However, if personality traits fluctuate over time, the correspondence between specific posts and a singular personality label may be inherently noisy, especially when treating each post, rather than each user, as the unit of analysis.

Future research could address these limitations by adopting weighted loss techniques to better handle class imbalance without discarding valuable data, and by designing hierarchical models that capture sequences of posts along a temporal axis. Incorporating multimodal data sources such as images and emojis could further enrich the representation of user expressions. In addition, developing ensemble frameworks that strategically integrate different classifier types, and conducting more comprehensive hyperparameter tuning, particularly for larger pretrained models like RoBERTa, could yield significant performance improvements. Finally, future work may benefit from embracing a dynamic modeling of personality, perhaps by framing the task as a time-series prediction problem or by exploring representations that allow for personality evolution over time.

7. Conclusion

In this project, we explored two modeling strategies for MBTI prediction based on users' text posts. One is direct 16-class classification, predicting the full personality type at once. The other is multi-binary classification, separately predicting the four MBTI dimensions: Energy (E/I), Information (S/N), Decision (T/F), and Lifestyle (J/P).

Starting from a logistic regression baseline model, we further explored alternative methods, including BERT, RoBERTa and T5. Although multi-binary classification and stronger pre-trained models like RoBERTa and T5 offered theoretical advantages, none outperformed BERT in terms of accuracy and stability. The multi-binary approach, while conceptually more aligned with MBTI's dimension structure, introduced compounded prediction errors and increased model complexity. RoBERTa showed promising expressive power through enhanced pretraining strategies but proved highly sensitive to hyperparameters and struggled with the dataset's nuanced language patterns. Meanwhile, T5's generative framework, despite offering flexibility, was less suited for discrete classification tasks and underperformed compared to discriminative models.

Following model selection, we successfully deployed an MBTI classification API to Hugging Face Spaces, enabling real-time public access to the model's capabilities. While the initial deployment demonstrates the model's potential, especially in directionally capturing personality traits, it also highlights areas for future enhancement.

Nevertheless, several limitations constrained the project's performance. These include reliance on undersampling techniques for class balance, the short length and isolated nature of text posts, lack of temporal modeling, and computational resource constraints that prevented extensive hyperparameter tuning and ensemble methods. Moreover, focusing exclusively on text without incorporating multimodal user data and assuming static personality types may have introduced additional noise.

Looking ahead, future research could focus on employing more sophisticated data balancing methods, capturing longitudinal user behaviors, integrating multimodal inputs, and adopting ensemble learning to combine the strengths of different architectures. Furthermore, rethinking personality prediction as a dynamic rather than static problem could open new perspectives for more accurate and meaningful modeling.

Overall, while our work establishes a strong baseline for MBTI personality prediction from text, it also reveals the challenges and opportunities inherent in this complex task. Continued refinement in both model design and data representation will be essential for achieving higher performance and deeper insights.

References

[1] MBTI Personality Type Twitter Dataset. Kaggle. https://www.kaggle.com/datasets/mazlumi/mbti-personality-type-twitter-dataset

[2] (MBTI) Myers-Briggs Personality Type Dataset. (2021, Feb 4). Alibaba Cloud. https://tianchi.aliyun.com/dataset/90272

GitHub Code and Dataset

https://github.com/your-username/your-repo-link