Statistical Learning
2025 Spring Project Report

Myers–Briggs Type Indicator(MBTI) Personality Classification and Prediction Using Statistical Learning

Longtian Shi 12210903

Kaixiang Liang 12212709

Yao Fu 12211653

1 Introduction

The Myers-Briggs Type Indicator (MBTI) is a well-known psychological framework that categorizes individuals into 16 distinct personality types based on four dichotomous axes: Introversion/Extraversion (I/E), Intuition/Sensing (N/S), Thinking/Feeling (T/F), and Judging/Perceiving (J/P). This project leverages statistical learning and natural language processing (NLP) techniques to explore the relationship between language patterns in social media text and MBTI personality types. We utilize a dataset of user-generated posts annotated with MBTI types to conduct exploratory data analysis, topic modeling (Latent Dirichlet Allocation), dimension reduction, and multi-label classification. Our goal is to develop a predictive model capable of classifying personality types from text content while addressing challenges such as severe class imbalance and the interpretability of linguistic features. The process includes text preprocessing, feature engineering, supervised classification (e.g., XGBoost, logistic regression, etc.), and validation using real-world text samples, ultimately resulting in a user interface for practical MBTI prediction.

2 Exploratory Data Analysis and Linear Discriminant Analysis

2.1 Dataset

The dataset used in this project includes users' text posts and their corresponding MBTI personality type labels. In terms of type distribution, the dataset contains the largest number of INFP type samples, followed by INFJ, while ESTJ type samples are the least numerous, indicating a certain degree of imbalance in the data. Additionally, we conducted a preliminary analysis of post length across different personality types. The results show that the post lengths of users across all types are mostly concentrated within the range of hundreds to thousands

of characters. This concentration trend provides a reasonable foundation for subsequent text cleaning and feature extraction.

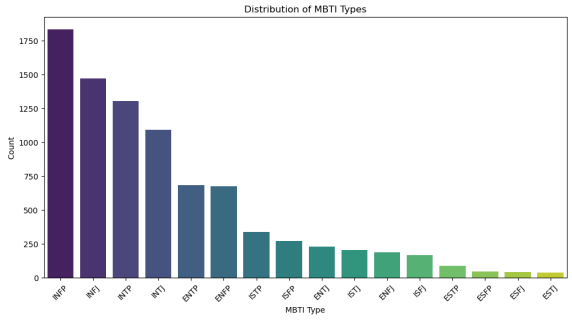


Figure 1: Histogram of sample counts across MBTI personality types

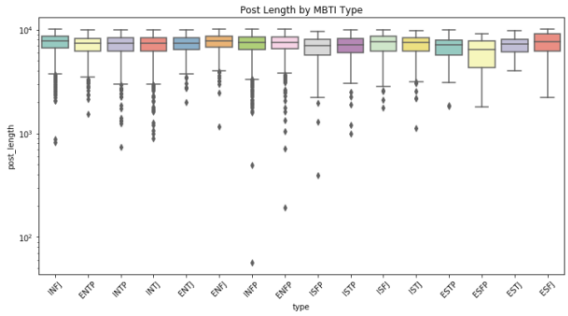


Figure 2: Boxplot of post lengths across different MBTI personality types

2.2 Sentiment Polarity

As shown in Figure 1, the dataset displays a notable imbalance in the distribution of MBTI personality types. INFP is the most prevalent type, followed by INFJ, while ESTJ appears least frequently. This skewed distribution should be taken into account in subsequent modeling to avoid potential bias. Furthermore, Figure 2 presents the distribution of post lengths across different MBTI types. The majority of posts fall within the range of several hundred to a few thousand characters, with relatively consistent patterns across personality types. This consistency in post length provides a reliable basis for downstream text preprocessing and feature extraction tasks.

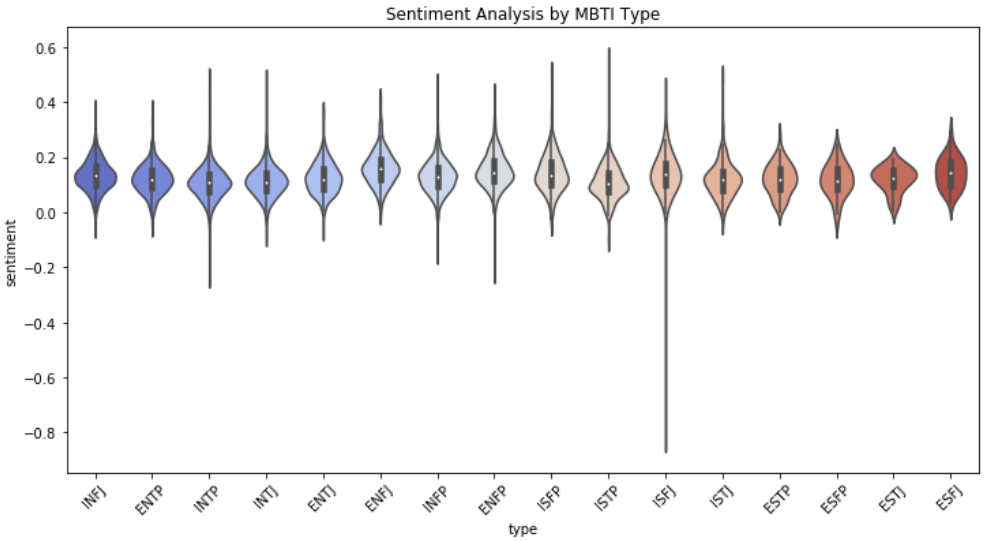


Figure 3: Violin plot of sentiment polarity in posts across MBTI personality types

2.3 Latent Dirichlet Allocation (LDA)

To uncover the underlying thematic structure in user posts, we employed an unsupervised topic modeling method—Latent Dirichlet Allocation (LDA). The primary objective was to explore whether users with different MBTI personality types tend to focus on different topics, thereby revealing the intrinsic connection between personality types and the thematic content of their language expressions. Before modeling, we performed comprehensive preprocessing on the text data. First, we uniformly converted the text to lowercase, removed URLs, HTML tags, and common abbreviations. Then, we used NLTK to tokenize the text and applied an expanded stopword list to remove meaningless filler words and MBTI type-specific vocabulary. To ensure the effectiveness of the topic model, we retained only high-information parts of speech such as nouns, verbs, and adjectives, and filtered out non-dictionary words and spelling errors through spell-checking. The cleaned and filtered text was converted into a bag-of-words representation as input for the LDA model.

Using the LDA model, we extracted four core themes that summarize the key content appearing in user comments and summarize the common interests and concerns of the entire sample.

Topic 0 centers on personal relationships and emotional experiences, as indicated by high-weight keywords such as "time," "love," "life," "friends," and "relationship". This topic likely captures users' reflections on interpersonal dynamics and affective states.

Topic 1 reflects a more abstract and socially oriented discourse, characterized by terms like "work," "world," "social," and "human".Comments associated with this topic may involve broader reflections on societal issues, values, and human behavior.

Topic 2 is closely tied to personality and self-analysis, with keywords including "personality," "test," "cognitive," and "introverted". This suggests that some users actively engage in discussing psychological constructs, personality theories, and their relevance to personal identity.

Topic 3 focuses more on everyday experiences, incorporating terms such as "school," "day," "music," "love," and "life". This theme appears to represent casual, daily-life narratives and emotional expressions, which may be particularly common among younger users.

Although the topics show some overlap, particularly between Topic 0 and Topic 3, their differences still offer useful distinctions. The relatively low coherence score (0.290) suggests that the topics are not highly distinct, which may be due to overlapping vocabulary or broad topic scopes. However, these topics provide a meaningful structure for understanding the general themes users focus on.

3 Dimensionality Reduction and Multi-label Classifier

In this part, we will discuss the dimensionality reduction and multi-label classifier.

3.1 Dimensionality Reduction

Ideally, if the dimensionality reduction is done successfully, the 16 personality types test clearly distinguishes into 16 groups. However, if we choose the optimal dimension as k=2 as the figure shown below, the actual situation is that the distribution of the 16 personality types is relatively concentrated and cannot be clearly distinguished though it still separate some

clusters.

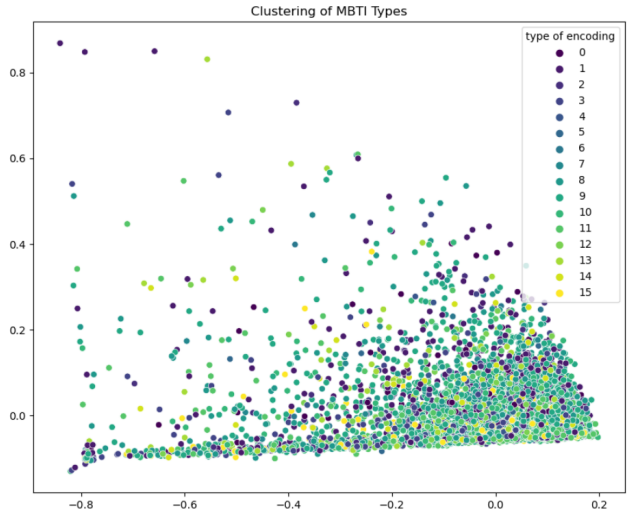


Figure 4: The PCA dimensionality reduction of choosing k=2

3.2 Multi-label Classifier

After that, we do the stratified cross-validation. we have to ensure that the distribution of MBTI types in each training/test set is the same as that of the entire set (for example: if INFP accounts for 20% in the data, then each set should have approximately 20% INFP).

```
=== 分层交叉验证结果 ===
平均准确率: 44.98% (±1.63%)
各折准确率: ['42.56%', '47.08%', '43.83%', '46.31%', '45.13%']
```

Figure 5: The CV result of the model

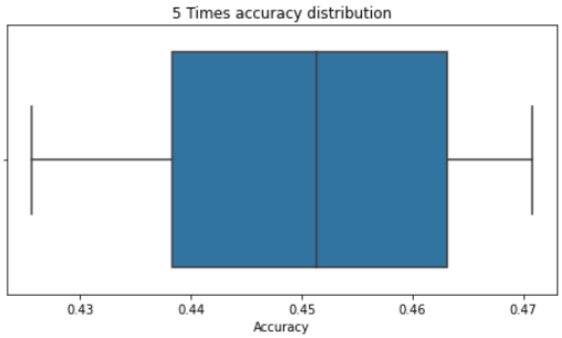


Figure 6: Boxplot of accuracy

we know from the figure that, for the accuracy rate, it is higher than random guessing the benchmark for 16 categories is 6.25%. Also, it has the low standard deviation (±1.63%), indicating that the model is relatively stable.

Also, from the current features have a certain predictive ability for the text, but the correlation between the features and the target is not strong enough. It can be initially used to screen out possible personality types (such as the coverage rate of predicting the top 3 types reaching 70%), but is not suitable for direct use in formal psychological assessment (it needs to be combined with other features or improved accuracy)

Next, we do the predicting extreme imbalance. For the High-performance types (ESFJ/ESFP/ESTJ), the F1-score is larger than 0.93 (almost perfect prediction). However, the sample size is extremely small (support less than 50), which may indicate overfitting. For the Low-performance types (ENFP/ENFJ), the F1-score is less than 0.6. While the sample size is relatively large (ENFP = 639), indicating that the real prediction is difficult.

Also, the recall rate is generally higher than the precision rate. The model tends to over-predict, labeling the samples as more types). For example, the ENFP which has large number of false positives and INFJ which is more accurate predictions but with more missed detections.

	precision	recall	f1-score	support
ENFJ	0.47	0.74	0.57	182
ENFP	0.36	0.72	0.48	639
ENTJ	0.53	0.77	0.63	225
ENTP	0.80	0.64	0.71	667
ESFJ	0.91	0.95	0.93	42
ESFP	1.00	0.91	0.96	47
ESTJ	0.95	0.97	0.96	39
ESTP	0.65	0.89	0.75	89
INFJ	0.86	0.66	0.75	1437
INFP	0.87	0.67	0.76	1778
INTJ	0.80	0.75	0.77	1075
INTP	0.83	0.73	0.77	1285
ISFJ	0.70	0.72	0.71	164
ISFP	0.63	0.72	0.67	267
ISTJ	0.64	0.81	0.71	202
ISTP	0.62	0.82	0.70	328
accuracy			0.71	8466
macro avg	0.73	0.78	0.74	8466
weighted avg	0.76	0.71	0.72	8466

Figure 7: The result of the CV

Then, we do the multi-label classifier. The data points represent the distribution of personalities: The sample is heavily skewed towards introverted (I) and intuitive (N) personalities.And the dimensions of thinking (T) and feeling (F) are relatively balanced. Also, the perceptive (P) type is more common than the judgmental (J) type.

Further, we can also find that sample is heavily skewed towards introverts possibily due to the anonymous online environment is more appealing to introverted individuals and the characteristics of the user group on the data collection platform. Next, the data is extremely biased towards intuitive types for most online text producers are abstract thinkers and S-type individuals may prefer offline expression. Relatively balanced still shows emotional types have a slight advantage since online expression is more emotional and clearly biased towards perceptive types, possibly because P-type individuals are more active in diverse online interactions.

```
=== MBTI各维度分布 ===
I_E: {0: 0.7720292936451689, 1: 0.22797070635483108}
N_S: {0: 0.860855185447673, 1: 0.13914481455232697}
T_F: {1: 0.5381526104417671, 0: 0.46184738955823296}
J_P: {1: 0.6024096385542169, 0: 0.39759036144578314}
```

Figure 8: The degree distribution of the plot

From the result, we know that The N/S dimension has the highest accuracy rate (87%). The J/P dimension has the lowest accuracy rate (about 0.72) (the distinction between Judging type (J) and Perceiving type (P) is the most challenging). The high accuracy of N/S dimension might be due to over-prediction of the majority type (N).

** I_E 维度 **

	precision	recall	f1-score	support
Not_I_E	0.87	0.87	0.87	1302
I_E	0.57	0.58	0.57	392
accuracy			0.80	1694
macro avg	0.72	0.72	0.72	1694
weighted avg	0.80	0.80	0.80	1694

** N_S 维度 **

	precision	recall	f1-score	support
Not_N_S	0.93	0.93	0.93	1483
N_S	0.49	0.50	0.50	211
accuracy			0.87	1694
macro avg	0.71	0.71	0.71	1694
weighted avg	0.87	0.87	0.87	1694

Figure 9: The multi-classifier result

** T_F 维度 **

	precision	recall	f1-score	support
Not_T_F	0.76	0.76	0.76	759
T_F	0.80	0.81	0.81	935
accuracy			0.79	1694
macro avg	0.78	0.78	0.78	1694
weighted avg	0.78	0.79	0.79	1694

** J_P 维度 **

	precision	recall	f1-score	support
Not_J_P	0.66	0.62	0.64	683
J_P	0.75	0.79	0.77	1011
accuracy			0.72	1694
macro avg	0.71	0.70	0.70	1694
weighted avg	0.72	0.72	0.72	1694

Figure 10: The multi-classifier result

For the visualization part, firstly, The I/E dimension (Extraverted vs. Introverted) has some obvious characteristic, the positive coefficient (Extraverted) shows some key words like explore, refreshing, light. While negative coefficient (Introverted) shows bored, personal, student. Also, for N/S Dimension (Intuition vs. Sensation), there also exists some key differentiating words. On the one hand, positive coefficient (Intuition) illustrates intuition, dark, remote; on the other hand, Negative coefficient (Sensation) shows step, edge, official. Last, for T/F Dimension (Thinking vs Feeling),the key characteristics is that positive Coefficient (Feeling) is about adore, idealistic, female and Negative Coefficient (Thinking) is about standard, repeat, android.

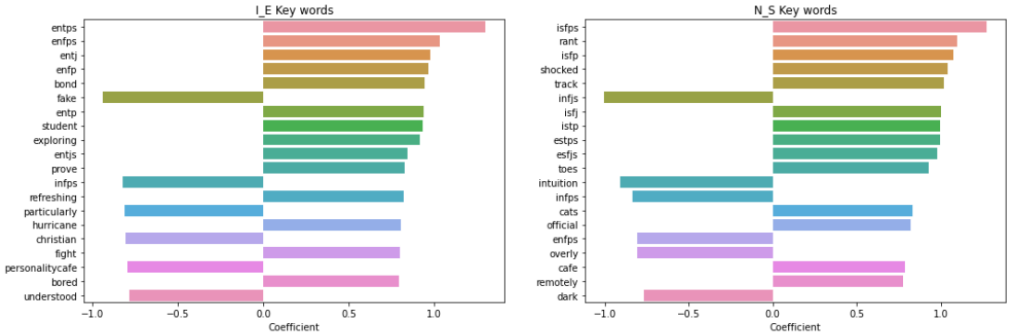


Figure 11: The Visualization

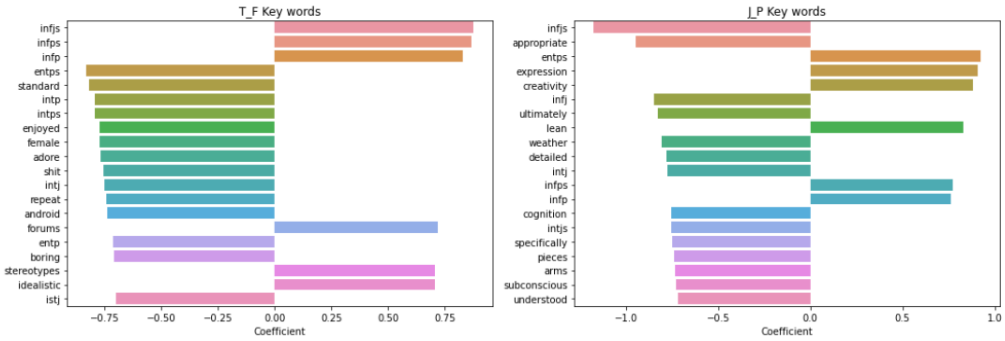


Figure 12: The Visualization

In conlusion, the linguistic pattern shows extroverted individuals tend to use more action verbs and positive emotion words. Introverted individuals tend to use more self-reflection-related words. For the cognitive differences, intuitive type prefers abstract concepts and metaphorical expressions. While sensory type tends towards concrete actions and formal language. For expression style, the emotional type contains more subjective evaluations and gender-related words. The thinking type contains more objective descriptions and technical terms

4 Natural Language Processing and Clustering

4.1 Overview of the NLP Pipeline

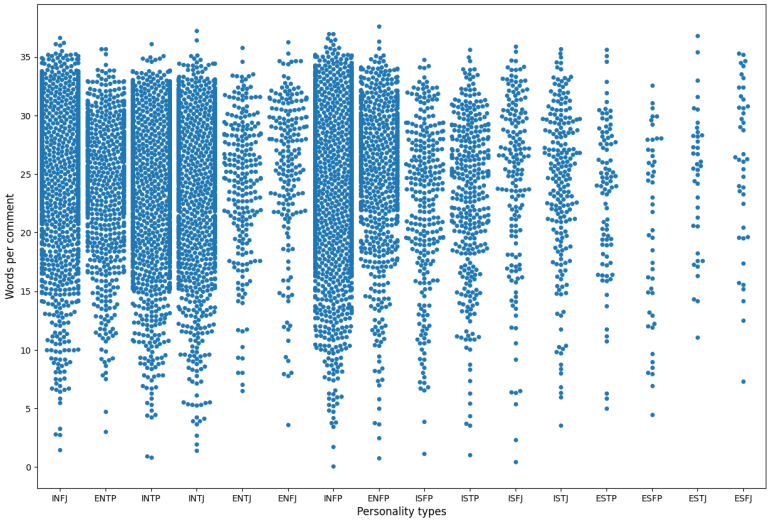
We first loaded the data of 8675 users, each with 50 concatenated posts and a 4-letter MBTI label. After confirming no missing values, we applied a custom text-cleaning function that (1) strips URLs and non-alphabetic characters, (2) inserts sentence-end markers, (3) lowercases, (4) removes MBTI type tokens to prevent label leakage, and (5) filters out posts with fewer than 15 words, yielding a cleaned corpus of 8466 entries, each tagged by its original MBTI.

4.2 Feature Extraction and Exploratory Analysis

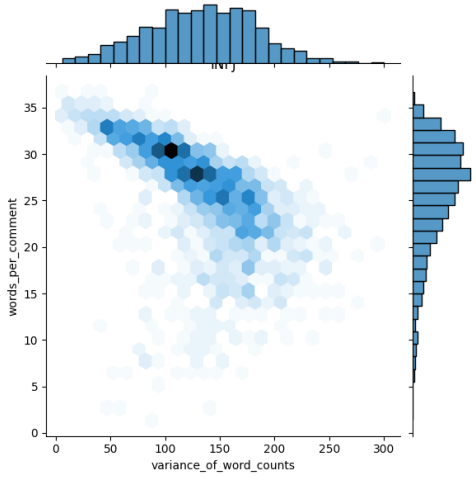
From each cleaned entry we computed two summary statistics—average words per comment and variance of words per comment across the 50 posts—then visualized their distribution

across the 16 MBTI types. Figure 13a shows a swarmplot of average words per comment by type, revealing that intuitive–thinking (NT) types tend to write longer comments, while sensing–feeling (SF) types show tighter clusters. Figure 13b presents a hex-bin joint plot of variance versus average word count for all users(we only show INFJ here due to the limit of the passage), highlighting that extroverts (E) have higher variance in comment length.

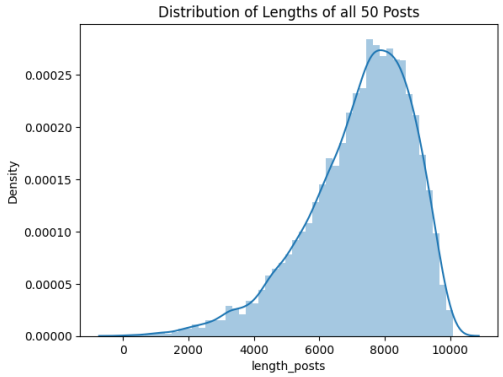
Next, we measured total post-length in characters and plotted its overall distribution (Figure 13c), finding a heavy right tail. A global word cloud (Figure 13d) surfaces the top tokens—"I," "to," "the," "a," "and"—which informed our subsequent stopword filtering.



(a) Average words per comment by MBTI type



(b) Hex-bin: variance vs. average words (INFJ example)



(c) Distribution of total post lengths

(d) Global word cloud of all posts

Figure 13: (a) Average words per comment; (b) Hex-bin joint plot; (c) Post-length distribution; (d) Global word cloud.

4.3 Vectorization and Baseline Classification

We converted the cleaned text into a document-term matrix using a count vectorizer limited to the 1000 most frequent tokens (min_df=0.1, max_df=0.7) and then applied TF–IDF

transformation. Splitting the data into 60-40 and 70-30 train/test sets (stratified by the 16-way MBTI label), we trained six baseline models (Random Forest, XGBoost, SGD, Logistic Regression, KNN, SVM). On the 70-30 split, XGBoost and Logistic Regression achieved the highest accuracies (57%), while KNN lagged (17%). Figure 14 displays these results.

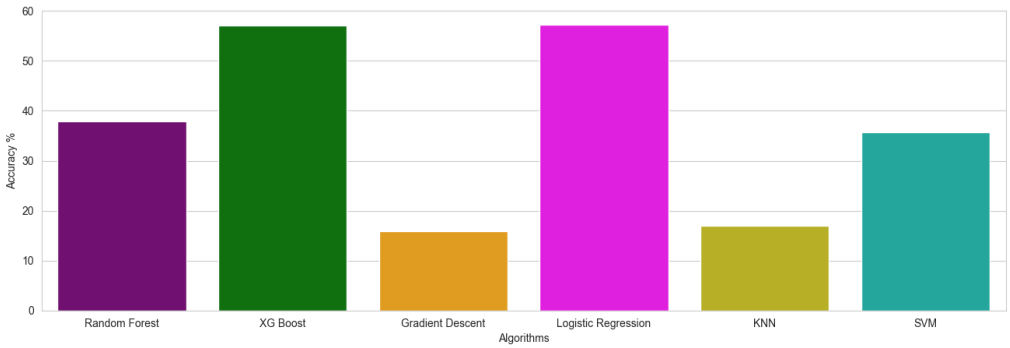


Figure 14: Accuracy of six classifiers on a 70-30 split

4.4 Four-Axis Binary Classification

To address the extreme class imbalance of 16-way prediction, we decomposed MBTI into four binary axes (I/E, N/S, T/F, J/P), encoding each label as 0/1. Using XGBoost with tuned hyperparameters (200 trees, max depth = 2, learning rate = 0.2) and a 33% test split, we trained four separate classifiers and obtained the accuracies shown in Table 1, confirming the efficacy of this decomposition.

Table 1: XGBoost accuracy by MBTI axis (33% test split)

Axis	Accuracy (%)
I/E: Introversion vs. Extraversion	77.40
N/S: Intuition vs. Sensing	85.89
T/F: Feeling vs. Thinking	69.79
J/P: Judging vs. Perceiving	63.95

4.5 External Validation and Extensions

We then demonstrated real-world applicability by predicting MBTI on three held-out texts (a cover letter, a poem, and a short essay). Each sample was preprocessed, vectorized, and passed through our four axis classifiers; their binarized outputs recombined into full types, yielding INFJ, INTP, and INTJ respectively (Table 2). Finally, we experimented with a soft-voting ensemble on the N/S axis—combining Logistic Regression, XGBoost, and Random Forest—which improved accuracy to 86.03%, and plotted a 5-fold learning curve to verify stability as training size increased.

Table 2: Predicted MBTI for three sample texts

Text Type	Excerpt	Predicted MBTI
Cover Letter	"I have a passion for teaching..."	INFJ
Poem	"They act like they care...suicide prevention."	INTP
Essay	"300 years...purpose...you're fucked."	INTJ

5 UI Designation

After completing model training, we built a user interface (UI) based on the trained XGBoost model to provide MBTI type prediction services. During the testing phase, we used DeepSeek to generate a self-introduction text that aligns with the INFP personality traits and input it into the interface for prediction. The model successfully output the INFP type, validating its predictive capability in real-world scenarios. To enhance user experience, we also integrated links corresponding to each MBTI type into the interface, enabling users to further explore their personality characteristics.

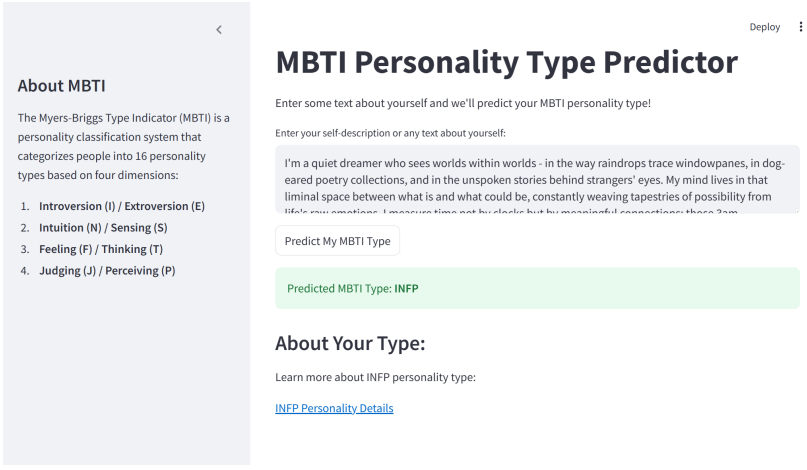


Figure 15: Main Interface of the MBTI Personality Type Prediction System

6 Conclusion and Discussion

This project systematically developed an MBTI personality prediction framework using statistical learning and NLP techniques. We began with exploratory analysis of 8,675 users' text posts, revealing intrinsic data characteristics: a skewed distribution toward INFP/INFJ types and consistent post lengths (hundreds to thousands of characters). Through Latent Dirichlet Allocation (LDA), we identified four core themes in user discourse—relationships, societal reflection, self-analysis, and daily experiences—establishing a link between personality types and topical engagement. For classification, we decomposed the 16-type problem into four binary axes (I/E, N/S, T/F, J/P), achieving peak accuracy on the N/S dimension (87%)

using XGBoost. Key linguistic patterns emerged: extroverts used action-driven vocabulary (explore, light), introverts favored self-referential terms (bored, student), and intuitive types employed abstract language (intuition, dark). The pipeline included rigorous text preprocessing, feature engineering (TF-IDF, post-length statistics), and stratified validation, culminating in a functional UI that successfully predicted INFP from synthetic text. This end-to-end workflow demonstrates the viability of leveraging textual data for personality insights.