

1 Introduction

With the recent renewal of the MBTI's popularity, posts or analyses about it have become widespread on public platforms. It seems that MBTI has become a way people use at first meetings.

The Myers-Briggs Type Indicator (MBTI) is a psychological framework primarily used to describe and measure personality traits by categorizing individuals into four dichotomous dimensions. In essence, it aims to reflect how people perceive the world and make decisions based on their internal experiences (Yang, 2022).

And the test is conducted through asking a series of questions, and the type of question is kind of like asking about the tendency to do something. Based on the responses, individuals receive a result represented by four letters, each corresponding to one of the four bipolar dimensions. According to Boyle (1995), these four dichotomous dimensions classify individuals as either extraverted (E) or introverted (I), sensing (S) or intuitive (N), thinking (T) or feeling (F), and judging (J) or perceiving (P). Combinations of the four preferences generate one of 16 personality types (e.g., ESFJ, ENFP, INTP, ISFJ), each associated with distinct behavioral tendencies, reflecting differences in attitudes, orientation, and decision-making styles. Also, on the official website, it will provide the future career options for you, like INFP, which may be more suitable for being an author. The percentage of each standard will also show in the final result.

The MBTI provides a widely recognized and accessible way to understand personality, making it a useful foundation for further behavioral and data-driven analysis.

2 Objective

1. Evaluate the reliability of the MBTI from a statistical perspective.
2. Explore the potential application of MBTI in social media behavior analysis.
3. Help people better understand personality traits and behavioral patterns.
4. Helping people eliminate stereotypes caused by MBTI personality types.

3 Problem Statements

1. Are the results of the MBTI personality test statistically robust and reliable?
2. Do the four dimensions of MBTI work independently, or are they connected in some way?
3. Do people with different MBTI types have different levels of activity on the internet?
4. Do significant differences exist in the interest preferences and behavioral patterns of different MBTI personality types on social networking sites?

4 Data Collection

1. “MBTI Personality Type Twitter Dataset”

Tweets were originally harvested from the public Twitter API by a third-party collector and later released on Kaggle by Mazlumi(<https://www.kaggle.com/datasets/mazlumi/mbti-personality-type-twitter-dataset>)

- 8,600 Twitter users, and 1 million raw tweets.
- Each user record contains the self-declared MBTI type (e.g., “ENFP”) taken from the user’s bio.
- The text has not been further cleaned or filtered—URLs, emojis, hashtags, and retweets remain. Researchers must therefore perform their preprocessing (tokenisation, stop-word removal, emoji handling, etc.) before analysis.

2. “KPMIRU Questionnaires Data”

Questionnaire responses compiled and shared on Kaggle by Pmenshih(<https://www.kaggle.com/datasets/pmenshih/kpmiru-questionnaires-data>)

- Contains every participant’s item-level answers to the full KPMIRU personality inventory (several dozen Likert-scale questions).
- Provides the scored results for all four MBTI dimensions—reported as continuous scores (0–100 per axis) as well as the final type label (e.g., “ISTJ”).

- Demographic fields (age range, gender, education) are included, enabling richer statistical controls.

Together, the Twitter dataset supplies large-scale, real-world language samples with self-reported types, while the KPMIRU dataset offers clean, psychometrically scored questionnaire data. The two sources complement each other for training and validating our emotion-aware MBTI models.

```
#reading the data sets
import pandas as pd
data1=pd.read_csv(r'E:\python\kpmi_ru_data.csv.zip')
print(data1)
```

	q1	t1	q2	t2	q3	t3	q4	t4	q5	t5	...	t63	e	i	s	n	t	\
0	1	8	2	8	2	8	1	8	1	8	...	8	32	18	16	19	26	
1	1	8	1	8	1	8	2	8	1	8	...	8	28	15	27	18	23	
2	1	8	1	8	1	8	1	8	1	8	...	8	25	15	25	16	19	
3	2	8	1	8	1	8	1	8	1	8	...	8	30	15	25	14	33	
4	2	8	2	8	1	8	1	8	1	8	...	8	22	22	23	18	33	
...	
99995	1	8	2	8	1	8	2	8	2	8	...	8	32	18	27	16	33	
99996	1	8	2	8	2	8	1	8	2	8	...	8	20	26	37	12	37	
99997	1	8	2	8	1	8	1	8	1	8	...	8	22	18	23	19	26	
99998	2	8	2	8	1	8	1	8	1	8	...	8	22	18	21	23	33	
99999	1	8	1	8	1	8	1	8	2	8	...	8	38	9	39	12	37	

	f	j	p	psychotype
0	24	19	44	ENTP
1	30	35	29	ESFJ
2	35	38	23	ESFJ
3	27	38	23	ESTJ
4	13	29	35	ISTP
...
99995	19	13	52	ESTP
99996	13	29	29	ISTP
99997	22	29	38	ESTP
99998	19	32	29	ENTJ
99999	22	45	23	ESTJ

[100000 rows x 133 columns]

Figure 1: reading data sets

```
import pandas as pd
data2=pd.read_csv(r"E:\python\twitter_MBTI.csv.zip")
print(data2)
```

	Unnamed: 0		text	label
0	0	@Pericles216	@HierBeforeTheAC @Sachinettiyl T...	intj
1	1	@Hispanthicckk	Being you makes you look cute ...	intj
2	2	@Alshymi	Les balles sont réelles et sont tirée...	intj
3	3		I'm like entp but idiotic Hey boy, do you wa...	intj
4	4	@kaeshurr1	Give it to @ZargarShanif ... He has...	intj
...
7806	7806	@sobsjjun	God,,pls take care 🙄 @sobsjjun Hir...	intp
7807	7807	@Ignis_02	wow last time i got intp https://t.c...	intp
7808	7808	@akupilled	A 100% @akupilled That SOMEONE wi...	entp
7809	7809		If you're #INTJ this one is for you What is ...	infj
7810	7810	@harry__lambert	@gucci hey can you dm me a pic...	istp

As illustrated in the two Figures above, the combined Kaggle sources provide information on:

1. Self-reported MBTI types for each respondent
2. Raw Twitter posts and basic tweet metadata linked to those MBTI labels
3. Demographic and psychometric questionnaire answers (KPMIRU survey)
4. Behavioral metrics such as posting frequency and topic keywords extracted from the tweets

5 MBTI in Measurement

5.1 Data Cleaning and Pre-processing

5.1.1 Handling Missing Values

1. We use `dropna()` to eliminate any records with missing or null entries. Fortunately, the dataset had no missing psychotype labels or scoring data.

```
# Delete null values
df_clean = df.dropna()
```

5.1.2 Outlier Detection and Correction

1. Outliers in numeric scores were identified using the IQR (interquartile range) method.
2. For each numeric column, we computed Q1, Q3. $IQR = Q3 - Q1$ represent the middle 50% of the data distribution. The lower_bound and upper_bound represent the boundaries of the normal range under the Interquartile Range (IQR) rule. And then we replaced values outside $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ with the column median.
3. The median substitution method mainly serves to stabilize the overall distribution and avoid the influence of extreme values.

```
# Select all columns in df_clean with numeric data types (either int or float) in preparation for outlier processing.
numeric_cols = df_clean.select_dtypes(include='number').columns

# Replace the outliers (utilizing the Interquartile Range (IQR) rule) with the median value of the column.
df_processed = df_clean.copy()

for col in numeric_cols:
    Q1 = df_processed[col].quantile(0.25)
    Q3 = df_processed[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    median_value = df_processed[col].median()

    # Replace outliers
    df_processed[col] = df_processed[col].apply(lambda x: median_value if x < lower_bound or x > upper_bound else x)
```

5.1.3 Export the cleaned data

After completing the outlier replacement and data cleaning process, we used `df.info()` to verify the integrity and structure of the cleaned dataset. The cleaned DataFrame was then exported using the `to_csv()` method, which saved it as `kpmi_ru_data(Cleaned).csv` for downstream analysis. The `index=False` parameter ensured that row indices were not written to the CSV file.

```
print(df_processed.info())
print(df_processed)
df_processed.to_csv(r'E:\python\kpmi_ru_data(Cleaned).csv', index = False)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Columns: 133 entries, q1 to psychotype
dtypes: float64(75), int64(57), object(1)
memory usage: 101.5+ MB
None
```

	q1	t1	q2	t2	q3	t3	q4	t4	q5	t5	...	t63	e	i	\
0	1	8.0	2	8.0	2	8.0	1	8.0	1	8.0	...	8.0	32.0	18.0	
1	1	8.0	1	8.0	1	8.0	2	8.0	1	8.0	...	8.0	28.0	15.0	
2	1	8.0	1	8.0	1	8.0	1	8.0	1	8.0	...	8.0	25.0	15.0	
3	2	8.0	1	8.0	1	8.0	1	8.0	1	8.0	...	8.0	30.0	15.0	
4	2	8.0	2	8.0	1	8.0	1	8.0	1	8.0	...	8.0	22.0	22.0	
...
99995	1	8.0	2	8.0	1	8.0	2	8.0	2	8.0	...	8.0	32.0	18.0	
99996	1	8.0	2	8.0	2	8.0	1	8.0	2	8.0	...	8.0	20.0	26.0	
99997	1	8.0	2	8.0	1	8.0	1	8.0	1	8.0	...	8.0	22.0	18.0	
99998	2	8.0	2	8.0	1	8.0	1	8.0	1	8.0	...	8.0	22.0	18.0	
99999	1	8.0	1	8.0	1	8.0	1	8.0	2	8.0	...	8.0	38.0	9.0	

	s	n	t	f	j	p	psychotype
0	16.0	19.0	26.0	24.0	19.0	44.0	ENTP
1	27.0	18.0	23.0	30.0	35.0	29.0	ESFJ
2	25.0	16.0	19.0	35.0	38.0	23.0	ESFJ
3	25.0	14.0	33.0	27.0	38.0	23.0	ESTJ
4	23.0	18.0	33.0	13.0	29.0	35.0	ISTP
...
99995	27.0	16.0	33.0	19.0	13.0	52.0	ESTP
99996	37.0	12.0	37.0	13.0	29.0	29.0	ISTP
99997	23.0	19.0	26.0	22.0	29.0	38.0	ESTP
99998	21.0	23.0	33.0	19.0	32.0	29.0	ENTJ
99999	39.0	12.0	37.0	22.0	45.0	23.0	ESTJ

[100000 rows x 133 columns]

The data in the figure 3.1 is the data after our data cleaning.

5.2 Model Building and Evaluation

Chi-square test is an on parametric statistical test to determine if the two or more classifications of the samples are independent or not Zibran, 2007. We all know that MBTI has four dimensions (Extraversion–Introversion (E–I), Sensing–Intuition (S–N), Thinking–Feeling (T–F) and Judging–Perceiving (J–P)). But whether or not these four dimensions interrelated or independent of each other stays unknown. For explanation, let’s consider the data presented in Figure 3.1 which comprising 100 000 respondents, providing information on the scoring fields of the four dimensions of MBTI. In order to find the answer, we use the chi-square test.

Each respondent’s type label (e.g., ENTP) was decomposed into its four constituent letters, and the sample was cross-classified into a $2 \times 2 \times 2 \times 2$ contingency table (16 cells) (shown in figure 3.2)

```
import pandas as pd
import numpy as np
from scipy.stats import chi2

# 1) 读入清理好的 CSV
FILE_PATH = r"E:\python\kpmi_ru_data(cleaned).csv"
df = pd.read_csv(FILE_PATH)

# 2) Extract the columns of four dimensions. (IE, SN, TF, JP)
mbti = df['psychotype'].astype(str).str.upper().dropna()
df_dims = pd.DataFrame({
    'IE': mbti.str[0], # E 或 I
    'SN': mbti.str[1], # S 或 N
    'TF': mbti.str[2], # T 或 F
    'JP': mbti.str[3] # J 或 P
})

# 3) Construct a 2x2x2x2 contingency table: Obtain the observed frequencies of 16 cells. counts[i,j,h,l]
import pandas as pd
import numpy as np

# Define dimension labels
index = pd.MultiIndex.from_product(
    [['E', 'I'], ['S', 'N'], ['T', 'F'], ['J', 'P']],
    names=['IE', 'SN', 'TF', 'JP']
)

# Build a DataFrame, using the values of counts as columns.
counts_df = pd.DataFrame({
    'Count': counts.flatten()
}, index=index)

print(counts_df.head(10))
```

IE	SN	TF	JP	Count
E	S	T	J	19038
			P	11168
		F	J	9506
			P	6604
	N	T	J	2933
			P	2333
		F	J	1914
			P	1802
I	S	T	J	14935
			P	8426

Then we calculate the marginal distribution (the distribution of each dimension separately), for example, $N_{IE} = [\text{count}(E), \text{count}(I)]$ (shown in figure3.3)

```
# Calculate marginal distribution and total number
N_IE = counts.sum(axis=(1,2,3))
N_SN = counts.sum(axis=(0,2,3))
N_TF = counts.sum(axis=(0,1,3))
N_JP = counts.sum(axis=(0,1,2))
N = counts.sum()

marginals = pd.DataFrame({
    'Dimension': ['IE', 'IE', 'SN', 'SN', 'TF', 'TF', 'JP', 'JP'],
    'Category': ['E', 'I', 'S', 'N', 'T', 'F', 'J', 'P'],
    'Count': np.concatenate([N_IE, N_SN, N_TF, N_JP])
})

print(marginals)
```

	Dimension	Category	Count
0	IE	E	55298
1	IE	I	44702
2	SN	S	81094
3	SN	N	18906
4	TF	T	64882
5	TF	F	35118
6	JP	J	60894
7	JP	P	39106

Under the null hypothesis of mutual independence, the expected frequency in each cell was computed as the product of the four one-dimensional marginal distributions multiplied by the sample size.(shown in figure3.4)

```
chi2_stat = ((counts - E)**2 / E).sum()
chi2_stat
```

```
np.float64(1342.93063615322)
```

Now, it's time to calculate the Pearson chi-square statistic, while O is observed value and E

is expected value.(shown in figure 3.5)

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

```
chi2_stat = ((counts - E)**2 / E).sum()
chi2_stat
```

```
np.float64(1342.93063615322)
```

Finally, with $\chi^2 = 1342.93$ and $df = 11$, we can compute the right - tailed probability corresponding to the chi - square statistic under 10 degrees of freedom, which is 0 (shown in figure 3.6). Because the p-value falls far below the conventional $\alpha = 0.05$ threshold, the null hypothesis is decisively rejected: the observed joint distribution of MBTI preferences deviates dramatically from what would be expected if the four indices varied independently. In practical terms, substantial associations exist among the E–I, S–N, T–F and J–P scales, corroborating earlier psychometric critiques that the MBTI dimensions are not orthogonal factors but overlap to a non-trivial extent. Actually, some researchers had also proven that the four dimensions of MBTI are not independent. Fleenor (1997) investigated the intercorrelations among MBTI continuous scores and found that while most dimension pairs demonstrated relatively low correlations, the correlation between the Sensing–Intuition (SN) and Judging–Perceiving (JP) scales was notably higher. Specifically, the study reported a correlation coefficient of $r = 0.41$ between SN and JP, indicating a moderate positive relationship. This finding has been replicated in other research and suggests that individuals who prefer intuition are more likely to also prefer perceiving. As such, the assumption of strict statistical independence between MBTI preference axes, particularly between SN and JP, may not hold, which is in line with the findings of our study.

```
df_val = 10
p_val = 1 - chi2.cdf(chi2_stat, df_val)
print(p_val)

0.0

if p_val < 0.05:
    print("Conclusion: Reject H0 — The four dimensions are not completely independent.")
else:
    print("Conclusion: H0 cannot be rejected — the four dimensions can be regarded as independent.")

Conclusion: Reject H0 — The four dimensions are not completely independent.
```

5.3 Exploratory Data Analysis (EDA)

5.3.1 MBTI Type Frequency

```
# 2) Frequency bar chart
plt.figure(figsize=(10,4))
df['psychotype'].value_counts().plot(kind='bar')
plt.title('Frequency of 16 MBTI Types')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

A frequency bar chart of the 16 MBTI types was generated using the `value_counts()` function on the `psychotype` column. The result was plotted using `matplotlib` and is shown in Figure 4.1.1.

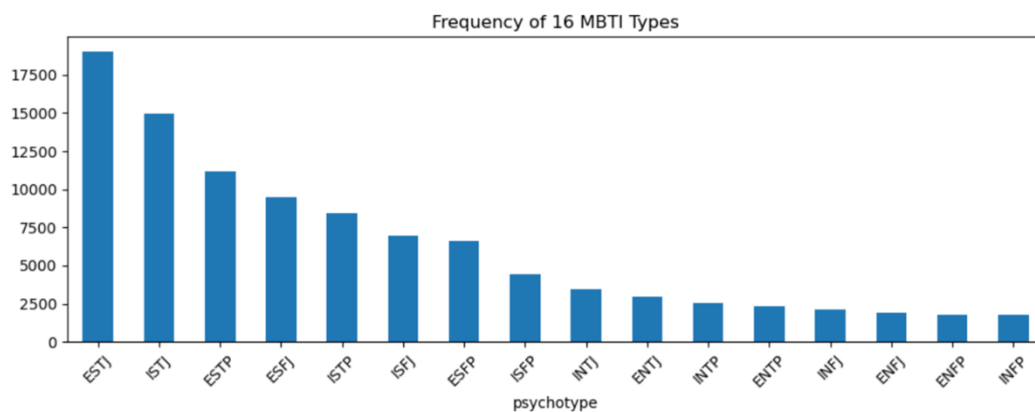


Figure 2: Frequency of 16 MBTI Types

As seen in Figure 4.1.1 above, the frequency of 16 MBTI types has been shown in a descending order. The type ESTJ ranks first, which shows that, based on our dataset, the proportion of ESTJ people is the highest. Conversely, the proportion of INFP people is the lowest. Additionally, it can be observed that the `_S_J` people rankings are relatively high in frequency, as the `_NF_` people rankings are relatively low.

Also, Figure 4.1.1 exhibits severe right oblique imbalance distribution, as the number of ESTJ people is nearly nine times as many as the number of INFP people. This phenomenon may be attributed to our database being originally from Kaggle, and certain MBTI type tends to participate in such investigations.

This contrasts with national MBTI distribution statistics reported in the MBTI® Manual, where ISFJ and ESFJ were found to be the most common types among U.S. adults (Myers et al., 1998), suggesting that our dataset, to a certain extent, fits the population-level trends. In distinct regions, the regional differences may influence MBTI type distributions in specific rankings.

5.3.2 Distribution of MBTI Dimension Scores

```
# 3) Histogram of scores in each dimension
score_cols = list('eisntfjp')
df[score_cols].hist(bins=20, figsize=(12,6))
plt.tight_layout()
plt.show()
```

Histograms of scores in each MBTI dimension were generated using pandas.DataFrame.hist and visualized with matplotlib as shown in Figure 4.1.2.

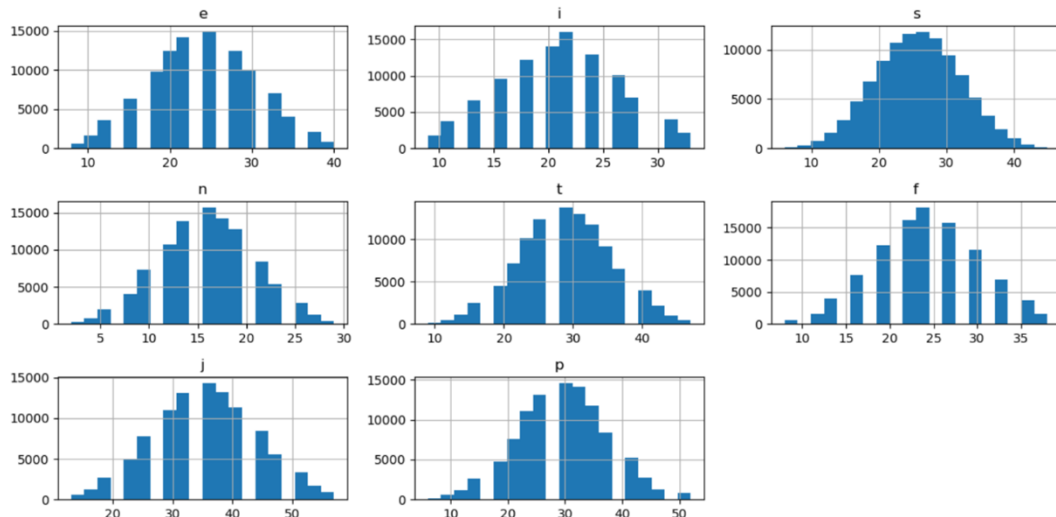


Figure 3: Histogram of MBTI Dimension Scores

As seen in Figure 4.1.2 above, from the score concentration trend, the eight MBTI poles vary in concentration. For example, E, S, and T scores are generally spread from 10 to 40, while N and I scores are more narrowly concentrated between 5 to 30. And in detail, in the

dimension of P and T, exhibit higher scores, which shows the preference toward Thinking and Perceiving. Compared to them, in the dimension of N and I, they exhibit lower scores, which indicates most people are more inclined toward Extraversion and Sensing.

From the data obtained from these diagrams, we can observe that most samples in these dimensions show distributional asymmetries, as not evenly distributed. Particularly for the high values of J and low values of N, this result matches the observation at 4.1.1 that S_J types dominate the database.

5.3.3 Descriptive Statistics of MBTI Dimensions

```
# 4) Describe statistics (mean, std, quartile, skewness, kurtosis)
stats = df[score_cols].describe().T
stats['skew'] = df[score_cols].skew()
stats['kurt'] = df[score_cols].kurt()
print(stats.round(2))
```

Descriptive statistics of each MBTI dimension were computed and consolidated into a single summary table, including skewness and kurtosis, to facilitate score distribution analysis shown in Table 4.1.1

	count	mean	std	min	25%	50%	75%	max	skew	kurt
e	100000.0	24.00	6.52	8.0	20.0	25.0	28.0	40.0	0.02	-0.38
i	100000.0	21.00	5.43	9.0	18.0	22.0	24.0	33.0	-0.06	-0.44
s	100000.0	26.19	6.65	6.0	21.0	27.0	31.0	45.0	-0.09	-0.19
n	100000.0	16.12	4.71	2.0	12.0	16.0	19.0	29.0	-0.11	-0.05
t	100000.0	28.89	6.64	9.0	23.0	28.0	33.0	47.0	0.00	-0.25
f	100000.0	24.05	5.92	8.0	19.0	24.0	27.0	38.0	-0.08	-0.25
j	100000.0	35.41	8.80	13.0	29.0	35.0	41.0	57.0	0.01	-0.31
p	100000.0	29.34	7.93	6.0	23.0	29.0	35.0	52.0	0.05	-0.19

Figure 4: Descriptive Statistics of MBTI Dimension Scores

Based on Table 4.1.3, it can be observed that the dimensions of I, S, N, and F showcase a notable positive skew distribution, as the dimension of N, with a value of -0.11, demonstrates left skewness most. This phenomenon tends to show that most people are more likely to be Sensing(S) and so on. And in contrast, the dimension of P with the value of +0.05 exhibits right skewness, implying a preference towards lower values, showing that Judging(J) dominates more in this database. Results of these also support the conclusion in Figure 4.1.1 that S_J people occupy the majority.

And from the kurtosis value, it can be seen that most values are around 0 and negative, which suggests that all the dimension is distributed platykurtic. With relatively concentrated values, this exhibits the loss of extreme outliers. Also, the maximum value and minimum value that all dimensions have can be used in the form of Max-Min, which demonstrates that J and P dimensions have the largest difference, indicating that individual differences between them are the strongest.

		S		N	
		T n(%)	F n(%)	T n(%)	F n(%)
E	J	ESTJ	ESFJ	ENTJ	ENFJ
		30 (12.2)	32 (13.1)	12 (4.9)	11 (4.5)
	P	ESTP	ESFP	ENTP	ENFP
		12 (4.9)	30 (12.2)	14 (5.7)	20 (8.2)
I	J	ISTJ	ISFJ	INTJ	INFJ
		29 (11.8)	21 (8.6)	2 (0.8)	2 (0.8)
	P	ISTP	ISFP	INTP	INFP
		12 (4.9)	10 (4.1)	3 (1.2)	5 (2.0)

Figure 5: MBTI Distribution of 16 Personality Types

Figure 4.1.3 MBTI Distribution of 16 Personality Types from Jang and Kim (2014), supporting the dominance of _S_J types observed in our dataset.

5.3.4 Correlation Analysis Between MBTI Dimensions

```
# 5) Correlation matrix heat map
plt.figure(figsize=(6,5))
sns.heatmap(df[score_cols].corr(), annot=True, cmap='coolwarm')
plt.title('Correlation between dimension scores')
plt.tight_layout()
plt.show()
```

A Pearson correlation heatmap was constructed to visualize linear relationships among MBTI dimension scores, as shown in Figure 4.1.4.

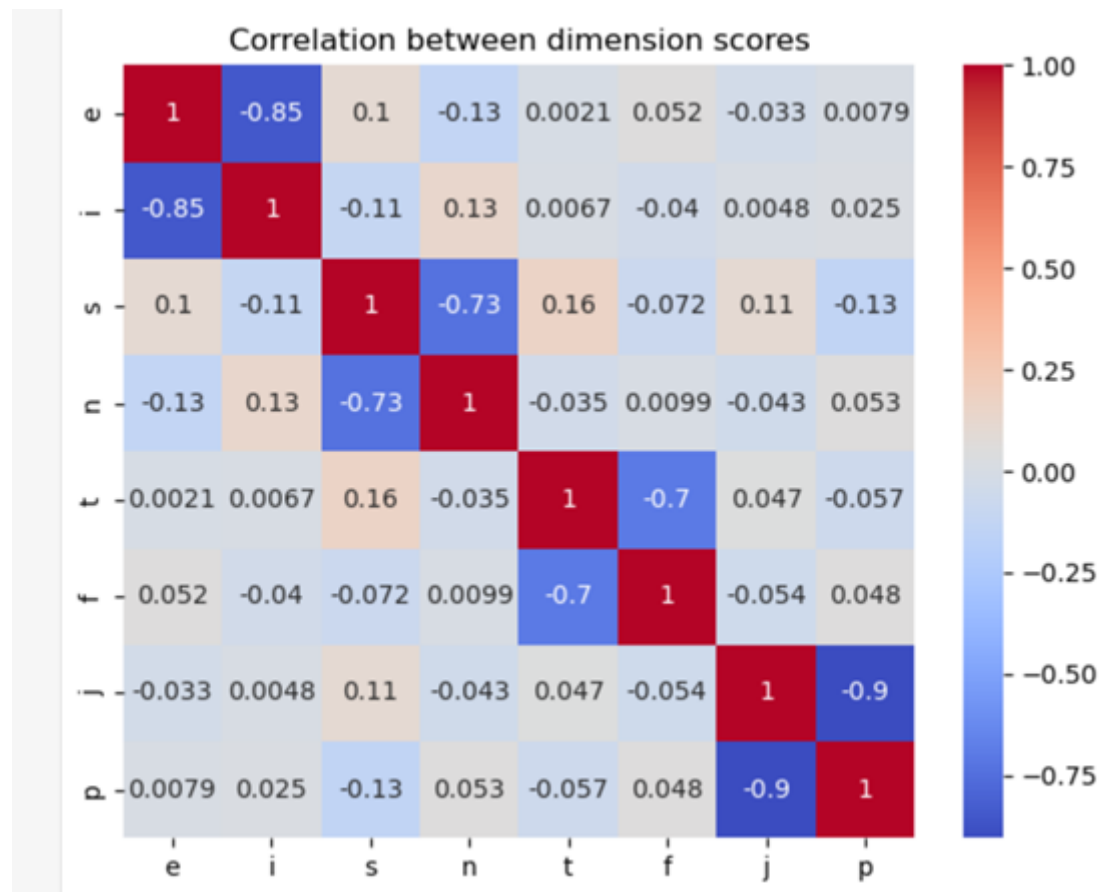


Figure 6: Correlation Heatmap of MBTI Dimension Scores

As seen in Figure 4.1.4 above, the correlation between the opposite dimensions, like E vs I, S vs N, T vs F, and J vs P. All of them are strongly negative, which is consistent with the design concept of MBTI, as the two poles of each dimension are opposed. As a result, the MBTI method has its rationality. This result aligns with (Li et al., 2024), who also observed that “the correlations between the four axes are generally weak, indicating that the personality traits on each axis are relatively independent” (p. 12).

Furthermore, in Figure 4.1.4, the correlation between other dimension pairs is low are nearly 0. This phenomenon supports that the dimensions are seemingly independent. While the correlation between S and T is +0.16, a bit larger than others, it indicates that someone who is Sensing(S) tends to be Thinking(T). In contrast, the correlation between S and P is -0.13, which indicates that those who are Sensing(S) tend to be Judging(J).

6 MBTI in Manifestation

In the previous section, we examined the statistical characteristics of MBTI. In this section, we turn to the manifestation level by investigating how different MBTI types vary in their topic preferences and expressive patterns on Twitter.

6.1 Topic Preferences of MBTI Personalities

To analyze the topic preference of different MBTI personalities, we adopt LDA modeling.

Latent Dirichlet Allocation(LDA) is a generative probabilistic model for uncovering hidden “topics” in a large collection of documents. It assumes each document is composed of a mixture of topics, and each topic is represented by a distribution over words. For example, a document might be 30% about “technology” and 70% about “health”, with each topic associated with its own common vocabulary.

In the LDA framework, the generation of a document is imagined as a two-step process:

1. a topic distribution is assigned to the document.
2. For each word in the document:
 - A topic is randomly chosen based on the document’s topic distribution.
 - A word is selected from the vocabulary of that topic.

During model training, this generative process is reversed: LDA infers the underlying topic structure based on the observed words in the documents. It estimates:

- the topic proportions for each document
- the most representative words for each topic

The specific LDA process of our project will be discussed in detail. But before we delve into it, we need to make sure our dataset is clean and ready for LDA modeling. Because LDA relies on word-frequency patterns to infer hidden topics, it’s important to remove noise, such as words that are meaningless or unrepresentative, so the model produces reliable topics.

6.1.1 Data Cleaning

6.1.2 LDA modeling

The cleaned data are organized by MBTI types, each as a tokenized and preprocessed text collection stored in a structured DataFrame. To perform LDA modeling, the input must be converted into a Bag-of-Words (BoW) corpus, where each document is represented as a list of (word_id, frequency) tuples. This requires mapping each unique word to a distinct integer ID, resulting in a dictionary that captures the vocabulary of the entire corpus. In this assignment, we apply filtering by removing words that appear in more than 20% of documents (no_above=0.2) and those that appear in fewer than 50 documents (no_below=50) to retain only representative terms.

```

1 import gensim.corpora as corpora
2 from gensim.models import CoherenceModel
3 def construct_initial_dict(source,no_above,no_below):
4
5     output = {T: {
6         "original_text": [],
7     } for T in MBTI_types}
8
9     output["all_original_text"]=[]
10    for T in tqdm(MBTI_types):
11        for i in source[T].data.index:
12            temp=source[T].data.loc[i,"posts"]
13            output[T]["original_text"].append(temp)
14            output["all_original_text"].extend(output[T]["original_text"])
15    output["overall_dict"]=corpora.Dictionary(output["all_original_text"])
16    output["overall_dict"].filter_extremes(no_above=no_above,no_below=no_below)
17    output["overall_dict"].compactify()
18    print("Size of dictionary:",len(output["overall_dict"]))
19    output["all_corpus"]=[output["overall_dict"].doc2bow(post_token) for post_token in output["all_original_text"]]
20    return output
21 initial_dict=construct_initial_dict(source=cleaned_data,
22                                   no_above=0.20,
23                                   no_below=50)
24 with open("Data/initial_dict.pkl",'wb') as f:
25     pickle.dump(initial_dict,f)
    
```

For a convenient inspection, we computed the overall term-frequency distribution by summing each token's bag-of-words counts across the entire corpus and stored the result table as a CSV file.

```

1 def check_corpus(corpus,dict,name=''):
2     result=pd.DataFrame(
3         [
4             list(range(len(dict))),
5             [0]*len(dict)
6         ]
7     ).T
8     result.columns=["word","frequency"]
9     for post in tqdm(corpus):
10        for word_tuple in post:
11            result.loc[word_tuple[0],"frequency"]+=word_tuple[1]
12    for i in result.index:
    
```



```

13     result.loc[i,"word"]=dict[i]
14     result=result.sort_values(by="frequency",ascending=False)
15     result.to_csv(f"Data/{name}id2word_result.csv")

```

Selecting an appropriate number of topics is crucial for LDA modeling. When the topic number is too small, unrelated content may be merged into the same topic, reducing interpretability. Conversely, an excessively large number of topics may fragment coherent semantic themes, leading to redundancy and overlap among topics. To determine the optimal number of topics for the LDA model, we trained multiple models with different numbers of topics. The semantic coherence and interpretability of the resulting topics were evaluated using the c_v coherence score, which measures how consistently related the top words within each topic are.

```

1  with open("Data/initial_dict.pkl","rb") as f:
2      initial_dict=pickle.load(f)
3  import gensim.corpora as corpora
4  import gensim
5  from gensim.models import LdaMulticore,CoherenceModel
6
7  def optimize_topic_num(
8      start,
9      end,
10     step,
11     dict=initial_dict["overall_dict"],
12     corpus=initial_dict["all_corpus"],
13     text=initial_dict["all_original_text"]
14 ):
15     output=pd.Series({},dtype=float)
16     topic_num_range=range(start, end+1, step)
17     for topic_num in tqdm(topic_num_range, desc="Calculating optimal topic number"):
18         # Train the LDA model (on all post data)
19         temp_lda_model=LdaMulticore(
20             corpus=corpus,      # Use the bag-of-words corpus of all posts
21             id2word=dict,      # Use the global dictionary
22             num_topics=topic_num,
23             random_state=100,
24             chunksize=100,      # Reduce chunksize to speed up update frequency
25             passes=10,          # Reduce passes to shorten total training time
26             iterations=50,      # Specify the number of iterations per pass
27             alpha=0.01,         # Use a smaller fixed value to promote topic sparsity
28             eta=0.01,           # Use a smaller fixed value to promote word sparsity
29             per_word_topics=False,
30             workers=None
31         )
32
33         # Evaluate the model
34         temp_chmodel=CoherenceModel(
35             model=temp_lda_model,
36             texts=text,
37             dictionary=dict,
38             coherence="c_v"
39         )
40         output[topic_num]=temp_chmodel.get_coherence()
41     print(output)

```

After a few trials, we find that 19 topics gives us the highest c_v coherence score.

count	16.000000
mean	0.113410
std	0.024352
min	0.076574
25%	0.093851
50%	0.110797
75%	0.128604
max	0.156459
Name:	compound, dtype: float64

Figure 7: Coherence scores for topic numbers 19 to 24. The highest score is observed at 19 topics, suggesting it as the optimal choice for this model.

We train the final LDA model with 19 topics using optimized settings, evaluate the model with `c_v` coherence score and save the model outputs, cleaned corpus, and topic descriptions for further analysis.

```

1  # Train the optimized final model with enhanced parameters for better convergence and topic separation
2  topics = 19
3
4  logging.basicConfig(level=logging.INFO, format='%(asctime)s - %(levelname)s - %(message)s')
5  # Optimized parameters for better convergence and topic distinction
6  lda_model = LdaMulticore(
7      corpus=initial_dict["all_corpus"],
8      id2word=initial_dict["overall_dict"],
9      num_topics=topics,
10     random_state=100,
11     chunksize=100,          # Significantly reduced: increases update frequency and improves convergence
12     passes=300,            # Moderately reduced: adjusted to work with other optimized parameters
13     iterations=150,        # Newly added: increase iterations per pass
14     alpha=0.01,           # Changed from 'asymmetric' to a small value: promotes document-topic sparsity and improves
                             # topic distinction
15     eta=0.01,             # Changed from 'auto' to a small value: promotes word-topic sparsity and reduces topic mixing
16     decay=0.5,            # Newly added: controls learning rate decay, improving convergence stability
17     offset=1.0,           # Newly added: initial value for learning rate
18     minimum_probability=0.01, # Newly added: filters out low-probability topic assignments
19     per_word_topics=False,
20     workers=None,         # Enables parallelization to speed up training
21     eval_every=20         # Reduces evaluation frequency to lower computational cost
22 )
23 # Model evaluation
24 chmodel = CoherenceModel(
25     model=lda_model,
26     texts=initial_dict["all_original_text"],
27     dictionary=initial_dict["overall_dict"],
28     coherence="c_v"
29 )
30 cv=chmodel.get_coherence()
31 cv
32 # Create unique directories for each LDA model
33 # That's because all variables are unique for each LDA model due to different stopword set
34 model_id=f"{topics}_{str(cv)[2:6]}"
35 # Model ID are designed as "[number of topics]_[CV score]", is unique for each model
36
37 path=f"output/lda_model/lda_{model_id}"
    
```

```

38 if not os.path.exists(path):
39     os.makedirs(path)
40
41 path=f"output/lda_model/lda_{model_id}/cleaned_data"
42 if not os.path.exists(path):
43     os.makedirs(path)
44
45 path=f"output/lda_model/lda_{model_id}/visualization"
46 if not os.path.exists(path):
47     os.makedirs(path)
48
49 # Save LDA model
50 with open(f"output/lda_model/lda_{model_id}/lda_{model_id}.pkl", 'wb') as f:
51     pickle.dump(lda_model, f)
52
53 # Save cleaned data
54 with open(f"output/lda_model/lda_{model_id}/cleaned_data/cleaned_data.pkl", "wb") as f:
55     pickle.dump(cleaned_data, f)
56
57 # Save original text
58 with open(f"output/lda_model/lda_{model_id}/all_original_text.pkl", "wb") as f:
59     pickle.dump(initial_dict["all_original_text"], f)
60
61 # Get all topics words and weights
62 all_topics_words = lda_model.show_topics(num_topics=-1, num_words=40, formatted=False)
63
64 markdown_content=f"## {topics} topics, cv={str(cv)[2:6]}\n\n"
65
66 for topic_id, topic_words_with_weights in all_topics_words:
67     markdown_content += f"### Topic {topic_id}:\n"
68
69
70     for word, weight in topic_words_with_weights:
71         markdown_content += f"- {word}: {weight:.4f}\n"
72     markdown_content += "\n"
73
74 # Save topic words and weights to markdown file
75 with open(f"output/lda_model/lda_{model_id}/lda_{model_id}.md", "w", encoding="utf-8") as f:
76     f.write(markdown_content)

```

After examining the model outputs, we observed that some meaningless or unrepresentative words still remain in the model. This is primarily because the Twitter dataset contains a substantial amount of slang that was not fully removed during the initial data cleaning process. To address this issue, we manually add these terms to the stopwords list and re-run the LDA modeling process iteratively until we obtain a model with a high c_v coherence score and minimal noise from irrelevant terms.

The model we finally settled on is model 19_5687, with the highest high c_v coherence score 0.5687. With this model, we will visualize the features of MBTI Twitter topics.

6.1.3 Result Visualization and Topic Evaluation

To start up the topic evaluation process, we need to import necessary libraries; define MBTI types, dimensions and groups; as well as load LDA model and cleaned data.

```

1 # Import necessary libraries

```

```

2 %matplotlib widget
3 import pyLDAvis
4 import pyLDAvis.gensim_models as gensimvis
5 import seaborn as sns
6 import matplotlib.pyplot as plt
7 from mpl_toolkits.mplot3d import Axes3D
8 import pandas as pd
9 import numpy as np
10 import pickle
11 from tqdm.auto import tqdm
12 import gensim.corpora as corpora
13 from gensim.models import LdaModel
14 from collections import defaultdict
15 from data_clean import Data_to_Clean,Data_to_Analyze
16 import warnings
17 warnings.filterwarnings('ignore')
18 from sklearn.cluster import KMeans
19 from sklearn.decomposition import PCA
20 from sklearn.preprocessing import StandardScaler
21
22 # Ensure the plots are displayed correctly
23 plt.rcParams['axes.unicode_minus']=False
24 sns.set_style("whitegrid")
25 sns.set_palette("husl")
26
27 # Define MBTI types
28 MBTI_types=[
29     'istj','isfj','infj','intj',
30     'istp','isfp','infp','intp',
31     'estp','esfp','enfp','entp',
32     'estj','esfj','enfj','entj'
33 ]
34
35 # Define MBTI dimensions
36 mbti_dimensions={
37     'E': ['estp','esfp','enfp','entp','estj','esfj','enfj','entj'],
38     'I': ['istj','isfj','infj','intj','istp','isfp','infp','intp'],
39     'S': ['istj','isfj','istp','isfp','estp','esfp','estj','esfj'],
40     'N': ['infj','intj','infp','intp','enfp','entp','enfj','entj'],
41     'T': ['intj','intp','entj','entp','istj','istp','estj','estp'],
42     'F': ['isfj','infj','isfp','infp','esfj','enfj','esfp','enfp'],
43     'J': ['istj','isfj','infj','intj','estj','esfj','enfj','entj'],
44     'P': ['istp','isfp','infp','intp','estp','esfp','enfp','entp']
45 }
46
47 # Define MBTI groups
48 mbti_groups={
49     "analysts":["intj","intp","entj","entp"],
50     "diplomats":["infj","infp","enfj","enfp"],
51     "sentinels":["istj","isfj","istp","isfp"],
52     "explorers":["isfp","istp","estp","esfp"]
53 }
54
55 # Load LDA model and data
56 # Model ID are designed as "[number of topics]_[CV score]",is unique for each model
57
58
59 def load_lda_data():
60     # Load LDA model
61     lda_model=pickle.load(open(f"output/lda_model/lda_{model_id}/lda_{model_id}.pkl","rb"))
62     print(f"Successfully loaded LDA model with {lda_model.num_topics} topics")
63
64     # Load original text data
65     all_original_text=pickle.load(open(f"output/lda_model/lda_{model_id}/all_original_text.pkl","rb"))
66     print(f"Successfully loaded original text data with {len(all_original_text)} documents")
67

```

```

68     return lda_model, all_original_text
69
70
71 # Load cleaned data grouped by MBTI types
72 def load_mbt_data():
73     file_path=f"output/lda_model/lda_{model_id}/cleaned_data/cleaned_data.pkl"
74     with open(file_path, 'rb') as f:
75         cleaned_data=pickle.load(f)
76     print(f"Cleand data loaded successfully")
77
78     return cleaned_data
79
80 # Execute file loading
81 lda_model, all_original_text=load_lda_data()
82 mbti_cleaned_data=load_mbt_data()
    
```

To better interpret topics and detect patterns in the data, we construct a dedicated class LDATopicAnalyzer, which integrates the LDA model with the MBTI-annotated dataset. This class performs several key functions: it transforms input texts into bag-of-words representations, computes topic distributions for each document, and aggregates these distributions by MBTI type to observe group-level thematic tendencies. It also supports generating interactive visualizations through pyLDAvis, which is created and saved to html. Furthermore, it provides tools to extract representative keywords per topic and optionally identify and exclude noise topics that carry little interpretive value.

```

1  # Create a class for LDA visualization
2  class LDATopicAnalyzer:
3      def __init__(self, lda_model, texts, mbti_data):
4          self.model=lda_model
5          self.texts=texts
6          self.mbt_data=mbti_data
7          self.dictionary=lda_model.id2word
8          self.corpus=[self.dictionary.doc2bow(text) for text in texts]
9          self.noise_topic=[]
10         self.topic_distributions=None
11         self.mbt_topic_distributions=None
12
13     def create_pyldavis_visualization(self, save_path=f"final_output/lda_visualization.html"):
14
15         print("Creating pyLDAvis visualization...")
16         # Prepare pyLDAvis visualization
17         vis_data=gensimvis.prepare(
18             self.model,
19             self.corpus,
20             self.dictionary,
21             sort_topics=False
22         )
23
24         # Save as HTML file
25         pyLDAvis.save_html(vis_data, save_path)
26         print(f"pyLDAvis visualization saved to: {save_path}")
27
28         # Display in notebook
29         pyLDAvis.enable_notebook()
30         return pyLDAvis.display(vis_data)
31
32     def get_topic_distributions(self):
33         """Get topic distributions for documents"""
34         print("Calculating topic distributions...")
    
```

```

35
36     topic_distributions=[]
37     for doc_bow in tqdm(self.corpus,desc="Calculating topic distributions"):
38         doc_topics=self.model.get_document_topics(doc_bow,minimum_probability=0)
39         topic_probs=[prob for _,prob in doc_topics]
40         topic_distributions.append(topic_probs)
41
42     self.topic_distributions=np.array(topic_distributions)
43     return self.topic_distributions
44
45     def calculate_mbt_topic_distributions(self):
46         """Calculate topic distributions for each MBTI type"""
47         print("Calculating topic distributions for each MBTI type...")
48
49         mbti_topic_dist={}
50
51         for mbti_type in MBTI_types:
52             if mbti_type in self.mbti_data and len(self.mbti_data[mbti_type].data) > 0:
53                 # Create corpus for documents of this MBTI type
54                 mbti_corpus=[self.dictionary.doc2bow(doc) for doc in self.mbti_data[mbti_type].data["posts"]]
55
56                 # Calculate topic distributions
57                 topic_sums=np.zeros(self.model.num_topics)
58                 doc_count=0
59
60                 for doc_bow in mbti_corpus:
61                     doc_topics=self.model.get_document_topics(doc_bow,minimum_probability=0)
62                     for topic_id,prob in doc_topics:
63                         topic_sums[topic_id]+=prob
64                     doc_count+=1
65
66                 # Calculate average topic distributions
67                 if doc_count>0:
68                     mbti_topic_dist[mbti_type]=topic_sums/doc_count
69                 else:
70                     mbti_topic_dist[mbti_type]=np.zeros(self.model.num_topics)
71             else:
72                 mbti_topic_dist[mbti_type]=np.zeros(self.model.num_topics)
73
74         self.mbti_topic_distributions=mbti_topic_dist
75         return mbti_topic_dist
76
77     def get_topic_words(self,num_words=10):
78         """Get keywords for each topic"""
79         topic_words={}
80
81         for topic_id in range(self.model.num_topics):
82             words=self.model.show_topic(topic_id,topn=num_words)
83             topic_words[topic_id]=[word for word,_ in words]
84         return topic_words
85
86     def add_noise_topics(self,*topic_ids):
87         """Define noise topics"""
88         for i in topic_ids:
89             self.noise_topic.append(i)
90
91 # Create analyzer instance
92 analyzer=LDATopicAnalyzer(lda_model,all_original_text,mbti_cleaned_data)
93 print("LDA topic analyzer created successfully!")
94
95 # Create pyLDAvis interactive visualization
96 vis=analyzer.create_pyldavis_visualization()
97 vis

```

The interactive topic visualization generated by pyLDAvis can be accessed at the following

link: https://dominicmin.github.io/Intro_to_DS_Assignment/lda_visualization.html. Figure 8 shows the initial state of the HTML file:

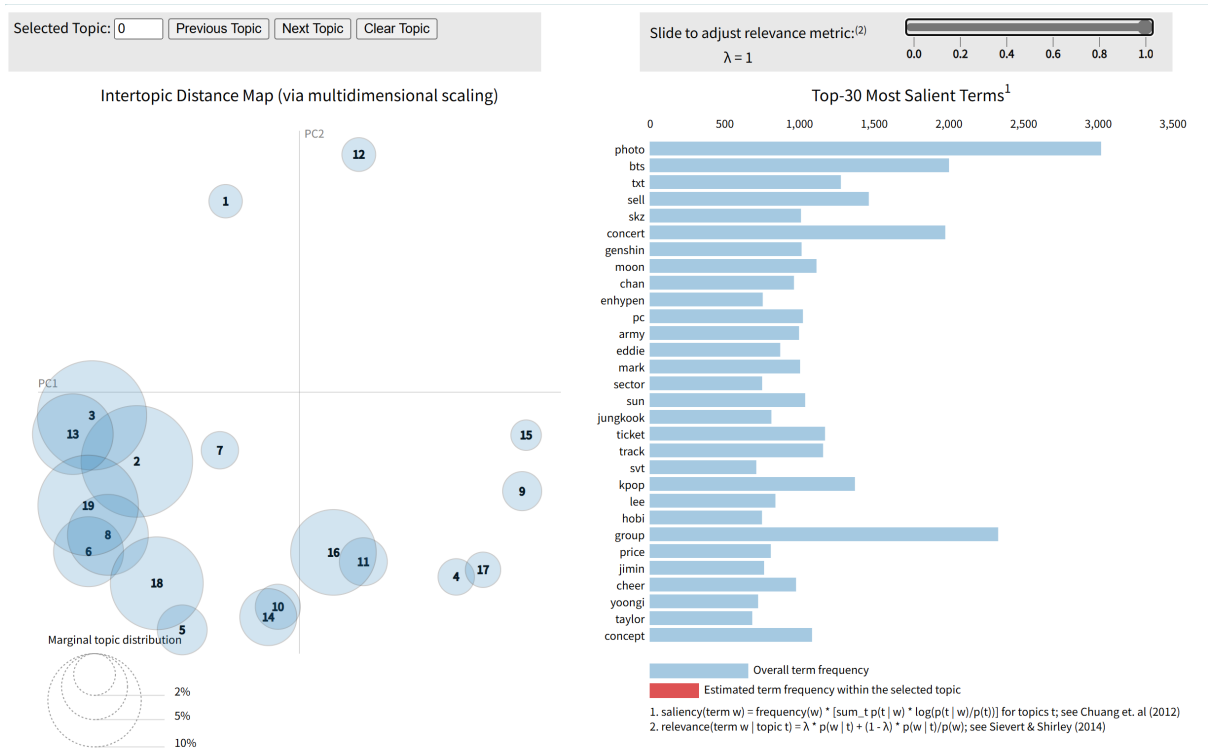


Figure 8: Screenshot of the initial state of the HTML file

Note: The topic indices in the HTML file are offset by +1 compared to the LDA model output (i.e., Topic 1 in HTML = Topic 0 in LDA).

As shown in Figure 8, on the left side of the HTML file, the Intertopic Distance Map visualizes the relationships between topics in two dimensions. A possible interpretation is that each bubble represents a distinct topic, with its area indicating the proportion of tokens (words) in the entire corpus attributed to that topic. Larger bubbles correspond to topics that account for a greater share of the corpus. The position of each bubble reflects semantic similarity—the farther apart two bubbles are, the less semantically related the corresponding topics. However, as will be demonstrated later, this interpretation fails to accurately characterize certain topics.

On the right side of the HTML page, the adjustable relevance metric of the selected topic is presented alongside the Top-30 Most Salient Terms. For each term, its overall frequency in the corpus is displayed as a blue histogram, while its estimated frequency within the selected topic is shown in red. Users can adjust the value of λ to observe changes in term saliency and the relationship between a term's overall frequency and its topic-specific relevance.

When λ is slid to the position 0, the terms that are unique in this topic are prior exhibited. These words are characterized by high distinctiveness, meaning they appear much more frequently in the selected topic compared to their frequency across the entire corpus. As a result, they serve as strong discriminators, helping to differentiate this topic from others. Although such terms may not be the most frequent within the topic itself, they often carry greater semantic specificity and are particularly useful for interpreting nuanced or domain-specific themes. However, because this setting emphasizes uniqueness over prevalence, it may occasionally highlight low-frequency or noisy terms, which should be interpreted with caution in topic labeling.

Meanwhile, when λ is slid to the position 1, these words represent the most commonly occurring terms in the topic and therefore reflect its core semantic content. This setting is particularly useful for understanding the dominant themes or main discourse of the topic. However, because it does not account for how exclusive a term is to the topic, some high-frequency terms that are also common in other topics may be included, potentially reducing the distinctiveness of the topic's representation.

From Figure 8, we can gain an overview of our modeling outcome. Topic 18 has the largest bubble, indicating that it is the most prevalent topic in the corpus. Topics 9, 15, 4, and 17 are located farther away from the other topics, suggesting a clear semantic distinction. In contrast, Topics 3, 13, 2, 8, 6, 19, and 18 exhibit notable spatial overlap, indicating only minor differences in their semantic content. In the whole corpus, the most reoccurring words are mostly related to the entertainment fandom and game fandom.

By selecting different topics and adjusting the value of λ , we can summarize the central theme of each topic, determine whether it is a noise topic, and assess its degree of semantic coherence.

Take topic 12 as an example, when λ is set to 0, the most relevant terms—such as trump, biden, republican, democrat, congress, and abortion—are highly distinctive and strongly associated with U.S. political discourse. These terms are not only topically specific but also exclusive to this topic, suggesting a focused and meaningful theme centered around American politics, government institutions, and social issues. In contrast, when λ is increased to 1, the top terms—such as state, country, child, gun, and law—shift toward higher-frequency words within the topic. Although some of these terms are more general, they still retain political relevance and semantic consistency with the topic's core, indicating that the theme is robust across different relevance metrics. In conclusion, the presence of consistently interpretable and con-

textually appropriate terms at both extremes of the λ scale demonstrates that topic 13 exhibits a relatively high degree of semantic coherence. The lack of function words, formatting artifacts, or off-topic vocabulary suggests that this is not a noise topic, but rather a well-defined and meaningful cluster within the corpus.

By applying this method to all the topic, we can conclude the following topic modeling result:

- High semantic coherence - 8 topics:

Topic 0: Astrology/Zodiac

Topic 3: K-pop Groups (TXT/NCT)

Topic 4: Western TV Shows (Stranger Things/Heartstopper)

Topic 6: Gaming (Genshin Impact)

Topic 8: K-pop Groups (SEVENTEEN/ATEEZ)

Topic 9: Western Pop Music/Celebrities

Topic 10: K-pop Group (BTS)

Topic 14: K-pop Group (ENHYPEN)

Topic 16: K-pop Group (Stray Kids)

- Medium semantic coherence - 6 topics:

Topic 2: Academic/Business/Personal Development

Topic 5: Gender Identity/LGBTQ+/Mental Health

Topic 7: Food/Weight Management/Eating Disorders

Topic 11: Merchandise Trading/Collectibles

Topic 12: Politics/Social Issues

Topic 17: Anime/Manga/Fan Culture

Topic 1: Daily Life and traveling

- Low semantic coherence - 1 topic:

Topic 15: K-pop Industry General

- Noise Topics - 2 topics:

Topic 13: Mixed Social Media Expressions

Topic 18: Generic Terms/Mixed Content

Interestingly, when linking the interpreted topics to their corresponding bubbles, we observe that although Topic 2 (Academic/Business/Personal Development) and Topic 12 (Politics/Social Issues) show considerable overlap in the Intertopic Distance Map, they are not semantically similar. This mismatch occurs because high-dimensional topic data is projected onto a two-dimensional space.

6.1.4 Topic Distribution

We begin by calculating and visualizing topic distribution.

```

1 # Calculate topic distributions for MBTI types
2
3 mbti_topic_dist=analyzer.calculate_mbti_topic_distributions()
4 topic_words=analyzer.get_topic_words()
5
6 # Create topic distribution DataFrame
7 topic_dist_df=pd.DataFrame(mbti_topic_dist).T
8 topic_dist_df.columns=[f"Topic {i}" for i in range(len(topic_dist_df.columns))]
9 topic_dist_df.drop([f"Topic {i}" for i in analyzer.noise_topic],
10                    axis=1,
11                    inplace=True)
12
13 print("MBTI type topic distribution calculation completed!")
14 print(f"Topic distribution matrix shape: {topic_dist_df.shape}")
15
16 # Display first few rows
17 print("\nPreview of topic distributions for each MBTI type:")
18 display(topic_dist_df.head())

```

```

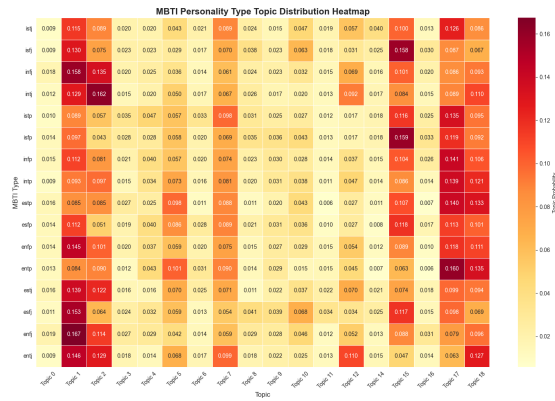
Calculating topic distributions for each MBTI type...
MBTI type topic distribution calculation completed!
Topic distribution matrix shape: (16, 18)

Preview of topic distributions for each MBTI type:
   Topic 0  Topic 1  Topic 2  Topic 3  Topic 4  Topic 5  Topic 6  Topic 7  Topic 8  Topic 9  Topic 10  Topic 11  Topic 12  Topic 14  Topic 15
intj 0.008851 0.115248 0.068992 0.020304 0.020248 0.043244 0.021340 0.088940 0.024472 0.015487 0.046669 0.018558 0.057390 0.040083 0.099723
isfj 0.009061 0.129999 0.074824 0.022828 0.023307 0.028510 0.017359 0.069671 0.038191 0.022679 0.063308 0.017913 0.031693 0.024531 0.157857
infj 0.017712 0.158324 0.134677 0.020251 0.025003 0.035576 0.014902 0.060592 0.023622 0.022519 0.037834 0.014991 0.068709 0.015630 0.100966
intp 0.011780 0.128567 0.161768 0.014841 0.020404 0.050019 0.016520 0.066708 0.026187 0.017296 0.020154 0.012905 0.092029 0.017059 0.084242
istp 0.009948 0.088690 0.056644 0.034672 0.046965 0.057443 0.033386 0.098363 0.030848 0.024917 0.027006 0.012163 0.017179 0.018405 0.116180

```

Figure 9: Preview of first few rows

A heatmap is generated based on the computed values to visually represent the results.



To assist with our interpretation with the heatmap, we calculate the overall popularity of each topic.

```

1 # Create topic keyword cloud summary
2 def create_topic_wordcloud_summary(topic_words,topic_dist_df):
3     """Create topic keyword summary"""
4     print("=" * 60)
5     print("Topic keyword summary")
6     print("=" * 60)
7
8     # Calculate overall popularity of each topic
9     topic_popularity=topic_dist_df.mean().sort_values(ascending=False)
10
11     for i,(topic_idx,popularity) in enumerate(topic_popularity.items()):
12         topic_num=int(topic_idx.replace('Topic',''))
13         print(f"\nTopic {topic_num} (Popularity: {popularity:.3f}):")
14         print(f"Keywords: {' '.join(topic_words[topic_num][:8])}")
15
16         # Find the MBTI type that most prefers this topic
17         topic_col=f"Topic {topic_num}"
18         if topic_col in topic_dist_df.columns:
19             top_mbtI=topic_dist_df[topic_col].nlargest(3)
20             print(f"Most preferred MBTI types: {' '.join([f'{mbti}({score:.3f})' for mbti,score in top_mbtI.items()])}")
21
22         if i >= 9: # Only show top 10 topics
23             break
24
25
26 create_topic_wordcloud_summary(topic_words,topic_dist_df)

```

This gives us the following summary:

```

=====
Topic keyword summary
=====

Topic 1 (Popularity: 0.122):
Keywords: photo,coffee,covid,trip,college,husband,felt,child
Most preferred MBTI types: enfj(0.167),infj(0.158),esfj(0.153)

Topic 17 (Popularity: 0.112):
Keywords: anime,chapter,fic,manga,holy,commission,au,arc
Most preferred MBTI types: entp(0.160),infp(0.141),estp(0.140)

Topic 18 (Popularity: 0.102):
Keywords: black,beat,star,war,holy,animal,ball,single
Most preferred MBTI types: entp(0.135),estp(0.133),entj(0.127)

Topic 15 (Popularity: 0.101):
Keywords: group,member,comeback,kpop,debut,dance,stalker,concert
Most preferred MBTI types: isfp(0.159),isfj(0.158),esfp(0.118)

Topic 2 (Popularity: 0.093):
Keywords: experience,self,human,study,thread,important,business,language
Most preferred MBTI types: intj(0.162),infj(0.135),entj(0.129)

Topic 7 (Popularity: 0.077):
Keywords: fast,weight,milk,fat,cream,ice,coffee,smell
Most preferred MBTI types: entj(0.099),istp(0.098),entp(0.090)

Topic 5 (Popularity: 0.061):
Keywords: trans,sex,gender,act,male,black,autistic,dick
Most preferred MBTI types: entp(0.101),estp(0.098),esfp(0.086)

Topic 12 (Popularity: 0.049):
Keywords: state,trump,country,child,law,gun,american,abortion
Most preferred MBTI types: entj(0.110),intj(0.092),estj(0.070)

Topic 10 (Popularity: 0.037):
Keywords: bts,army,jungkook,jimin,hobi,yoongi,taehyung,tae
Most preferred MBTI types: esfj(0.068),isfj(0.063),istj(0.047)

Topic 4 (Popularity: 0.030):
Keywords: eddie,steve,mike,stranger,robin,max,strange,el
Most preferred MBTI types: istp(0.047),entp(0.043),infp(0.040)

```

With the aid of the summary statistics, we interpret the heatmap and draw conclusions from two analytical perspectives.

Analyzing from the perspective of individual topics, we can find distinct MBTI engagement patterns across different semantic domains.

- Topic 15 (K-pop fandom activities): characterized by group debuts, comebacks, and concerts; demonstrates strong engagement among ISFP, ISFJ, and ESFP types, underscoring the intersection of sensory enjoyment and emotional engagement.
- Topic 14 (lightweight, emotionally-driven daily expression): resonates with SP types—particularly SFPs and STPs—who prioritize spontaneity and immediate experiential sharing.
- Topic 7 (sensory experiences related to food and comfort): emphasizes fast food, coffee, and cream, aligning with ISTP and ESTP preferences for direct sensory satisfaction.
- Topic 17 (strategic, abstract, and fandom-oriented discussions): involves anime, manga, and fan fiction; attracts ENTP, INFP, and ESTP types, indicating alignment with analytical and imaginative discourse.

- Topic 2 (experiential and intellectual reflections): appeals to INTJ, ENTJ, and INFJ types, reflecting their preference for profound, thoughtful, and self-reflective content.
- Topic 18 (adventurous and competitive themes): includes sports and competitions; engages ENTP, ESTP, and ENTJ personalities drawn to intense experiences.
- Topic 5 (identity and socially charged discussions): covers gender, sexuality, and societal roles; garners interest from ENTP, ESTP, and ESFP types open to explorative and provocative topics.
- Topic 12 (politically charged discourse): focuses on societal debates like Trump, guns, and abortion; attracts ENTJ, INTJ, and ESTJ types inclined toward policy and governance.
- Topic 10 (BTS fandom): features BTS member names; resonates with ESFJ and ISFJ types, highlighting community-oriented engagement.
- Topic 4 (Stranger Things franchise): references characters like Eddie and Steve; shows engagement from ISTP, ENTP, and INFP types, indicating analytical and introspective affinities.

Analyzing from the perspective of MBTI functional groups, clearer behavioral and cognitive preferences emerge.

- NF types consistently gravitate toward emotionally and symbolically rich discussions (Topics 1, 2), emphasizing their values-driven communication style and expressive tendencies.
- SP types distinctly favor immediate experiential and sensory-focused topics (Topics 7, 14, 15), indicative of their present-oriented, sensory-driven inclinations.
- NT types predominantly engage in abstract reasoning, strategic analyses, and fandom discussions that encourage complex intellectual engagement (Topics 2, 5, 17).
- SJ types display balanced interests, engaging prominently with mainstream cultural phenomena and structured, community-centric discussions (Topics 1, 10, 17), reflecting their preference for structured interaction, practical feasibility, and communal harmony.

6.1.5 Topic Preference between MBTI Dimensions and Groups

6.2 Expressive Patterns of MBTI Personalities

6.2.1 Data Cleaning and Pre-processing

Expression pattern analysis shares the same initial steps as LDA modeling: removing URLs, emojis, and tags. Please refer to the LDA data-cleaning section for details. We only need to load the cleaned data for our expressive pattern analysis.

```

1 from data_clean import Data_to_Clean, Data_to_Analyze
2 import pickle
3 import os
4 import pandas as pd
5 import copy
6 import json
7 import logging
8 from tqdm.auto import tqdm
9 import matplotlib.pyplot as plt
10 tqdm.pandas()
11
12 MBTI_types=[
13     'istj', 'isfj', 'infj', 'intj',
14     'istp', 'isfp', 'infp', 'intp',
15     'estp', 'esfp', 'enfp', 'entp',
16     'estj', 'esfj', 'enfj', 'entj'
17 ]
18 cleaned_data={T:None for T in MBTI_types}
19
20 for type in cleaned_data.keys():
21     file_path=f"Data\\cleaned_data\\{type}_cleaned.pkl"
22     try:
23         with open(file_path, 'rb') as f:
24             cleaned_data[type] = pickle.load(f)
25     except FileNotFoundError:
26         print(f"Error: File not found at {file_path}")
27     except pickle.UnpicklingError:
28         print(f"Error: Could not unpickle the file {file_path}. It might be corrupted or not a valid pickle file.")
29     except Exception as e:
30         print(f"An unexpected error occurred: {e}")

```

To check whether we have successfully imported the cleaned dataset, we print out the 0th column, "posts" row of infp.

```

1 infp=cleaned_data["infp"]
2 infp.data.loc[0,"posts"]

```

From the output, we can see that there is no URLs, emojis, and tags in our data, which means that we have successfully cleaned and read from the cleaned dataset. Our data is ready for analysis.

```
[50]: ['forest',  
      'elizabeth',  
      'holly',  
      'fuss',  
      'seat',  
      'pant',  
      'best',  
      'friend',  
      'day',  
      'yes',  
      'decoration',  
      'old',  
      'day',  
      'change',  
      'present',  
      'guess',
```

Figure 10: Cleaned Data printed out

6.2.2 Linguistic Style Patterns of MBTI Types

We select 4 distinct linguistic dimensions to represent linguistic styles:

1. **Sentence Quantity:** A larger sentence quantity suggests that the user tends to express themselves in shorter, more frequent statements. This may reflect a conversational, spontaneous communication style, often associated with extraversion or emotional openness.
2. **Word Count:** Higher word count may indicate a more elaborative or expressive communication habit, suggesting the user tends to provide more context or detail. This could reflect traits such as thoughtfulness, emotional depth, or even analytical thinking.
3. **Upper-Case Ratio:** A higher ratio of uppercase letters often signifies stronger emotional

expression, emphasis, or a more dramatic tone. It may reveal higher emotional arousal or a more assertive communication style, which could be linked to personality traits like enthusiasm or dominance.

4. Reading Ease(reflected by both Flesch and GF index):Higher reading ease scores imply simpler sentence structures and more familiar vocabulary, often seen in casual or informal communication. In contrast, lower scores suggest complexity and abstractness, which could indicate intellectual engagement, formality, or introversion. In our project, we adopted two kinds of rating systems to evaluate the reading ease of a context. The higher the Flesch Reading Ease score is, the easier the context can be understood; the lower the GF index is, the easier the context can be understood.

These values are first computed before further analysis.

1. Class Definition & Constructor

```

1  class Data_to_Analyze(Data_to_Clean):
2      def __init__(self, type, source=raw_data):
3          super().__init__(source)
4          self.data = self.data.loc[self.data["type"] == type].reset_index(drop=True)
5          self.data_to_vec = None
6          self.basic_identities = pd.Series({
7              "type": type,
8              # Number of sentences in a post
9              "sentence_quantity": [],
10             "ave_sentence_quantity": None,
11             # Number of words in a post
12             "word_count": [],
13             "ave_word_count": None,
14             # Ratio of upper case characters in a post
15             "upper_ratio": [],
16             "ave_upper_ratio": None,
17             # Two readability indicators: Flesch Reading Ease and Gunning Fog Index
18             "reading_ease": [],
19             "ave_reading_ease": None,
20             "GF_index": [],
21             "ave_GF_index": None,
22             # Overall sentiment indicator (VADER)
23             "overall_vader_score": None
24         })
    
```

Defines a class Data_to_Analyze that inherits from Data_to_Clean.

- **Data subset** – The constructor calls `super().__init__(source)` to initialize the parent class, then filters the dataset so that only rows whose type column equals the target MBTI type remain (`self.data`).
- **Feature container** – `basic_identities` is a `pd.Series` used to store a collection of text statistics: sentence count, word count, upper case ratio, readability metrics (Flesch Reading Ease and Gunning Fog Index) and an overall VADER sentiment score.

2. Sentence Count

```
1 def get_sentence_quantity(self):
2     for post in self.data["posts"].values:
3         self.basic_identities["sentence_quantity"].append(len(post))
4     self.basic_identities["ave_sentence_quantity"] = ave(self.basic_identities["sentence_quantity"])
```

Iterates through every post (each post is already a list of sentences). The length of the list gives the number of sentences, which is appended to sentence_quantity. Finally, the helper ave() computes their mean.

3. Word Count

```
1 def get_word_count(self):
2     for post in self.data["posts"].values:
3         total = 0
4         for sentence in post:
5             total += len(sentence.split(" "))
6         self.basic_identities["word_count"].append(total)
7     self.basic_identities["ave_word_count"] = ave(self.basic_identities["word_count"])
```

For each post, it splits every sentence on whitespace to count words and sums them into total. The total per post is stored in word_count, and the average is calculated afterwards.

4. Upper Case Character Ratio

```
1 def get_upper_ratio(self):
2     for post in self.data["posts"].values:
3         char_count = 0
4         upper_count = 0
5         for sentence in post:
6             for ch in sentence:
7                 if ch.isalpha():
8                     char_count += 1
9                 if ch.isupper():
10                    upper_count += 1
11         if char_count:
12             self.basic_identities["upper_ratio"].append(upper_count / char_count)
13     self.basic_identities["ave_upper_ratio"] = ave(self.basic_identities["upper_ratio"])
```

Traverses every character of every sentence, counting alphabetic characters (char_count) and, among them, the upper case ones (upper_count). The ratio per post is stored; the overall mean is then computed.

5. Readability Metrics

```
1 def get_readability(self):
2     reading_ease = []
3     GF_idx = []
4     for post in self.data["posts"].values:
5         concatenated = post[0]
6         for idx in range(1, len(post)):
7             concatenated += post[idx]
8         reading_ease.append(textstat.flesch_reading_ease(concatenated))
9         GF_idx.append(textstat.gunning_fog(concatenated))
10    self.basic_identities["reading_ease"] = reading_ease
```

```

11 self.basic_identities["ave_reading_ease"] = ave(reading_ease)
12 self.basic_identities["GF_index"] = GF_idx
13 self.basic_identities["ave_GF_index"] = ave(GF_idx)

```

Each post's sentences are concatenated into a single string, after which textstat computes Flesch Reading Ease and Gunning Fog Index. Both per post values and their averages are stored.

The calculation results are printed and displayed.

```

1 mbti_identities={
2     T.upper():{
3         k:cleaned_data[T].basic_identities[k]
4         for k in [
5             "ave_sentence_quantity",
6             "ave_word_count",
7             "ave_upper_ratio",
8             "ave_reading_ease",
9             "ave_GF_index"
10        ]
11    }
12    for T in MBTI_types
13 }
14 mbti_identities=pd.DataFrame(mbti_identities).T
15 mbti_identities

```

	ave_sentence_quantity	ave_word_count	ave_upper_ratio	ave_reading_ease	ave_GF_index
ISTJ	140.922780	1584.861004	0.125113	53.508301	16.561120
ISFJ	139.890110	1542.766484	0.139261	58.487995	14.762555
INFJ	136.120151	1662.837275	0.108945	59.505421	14.098761
INTJ	142.160051	1768.897567	0.111903	55.228515	15.652458
ISTP	142.599388	1439.327217	0.135959	50.868624	18.008930
ISFP	138.239782	1468.850136	0.147333	51.682698	17.722262
INFP	135.311232	1538.657566	0.133015	54.009626	16.649938
INTP	138.747226	1557.081381	0.124319	50.273724	18.050296
ESTP	138.510000	1528.890000	0.144286	54.100900	16.754100
ESFP	147.137931	1500.454023	0.157108	54.208218	16.571782
ENFP	141.171468	1655.497942	0.136338	55.011289	16.238724
ENTP	147.183709	1617.429809	0.125773	48.465754	18.679307
ESTJ	150.148148	1861.827160	0.099657	55.780123	15.963333
ESFJ	139.666667	1571.104762	0.141718	58.292857	15.020381
ENFJ	141.915058	1690.857143	0.119560	56.535425	15.473456
ENTJ	142.960573	1713.476703	0.116676	55.087849	15.837384

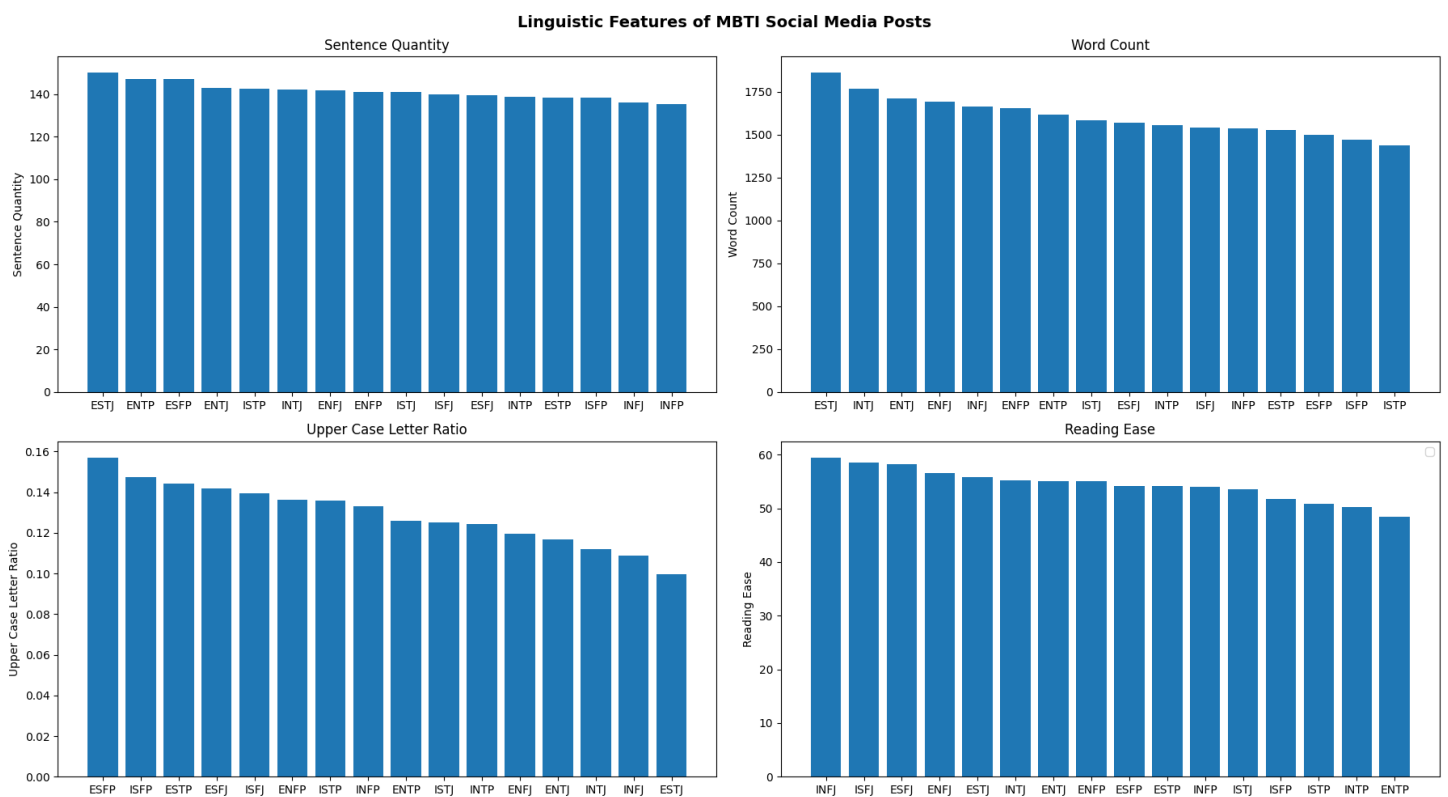
Figure 11: Table of Linguistic Style Features of MBTI Social Media Posts

In order to better interpret the underlying patterns in the table, we present descending histograms for visual analysis.

```

1 column_mapping={
2     "ave_sentence_quantity":"Sentence Quantity",
3     "ave_word_count":"Word Count",
4     "ave_upper_ratio":"Upper Case Letter Ratio",
5     "ave_reading_ease":"Reading Ease"
6 }
7 fig,axes=plt.subplots(2,2,figsize=(18,10))
8 axes=axes.flatten()
9 for i,col in enumerate(mbti_identities.columns[:4]):
10     axes[i].bar(mbti_identities[col].sort_values(ascending=False).index, mbti_identities[col].sort_values(ascending=False))
11     axes[i].set_ylabel(f"{column_mapping[col]}")
12     axes[i].set_title(f"{column_mapping[col]}")
13 fig.suptitle("Linguistic Features of MBTI Social Media Posts",fontsize=14, fontweight='bold')
14 plt.legend()
15 plt.tight_layout()
16 plt.savefig("final_output/ling_features.png")
17 plt.show()

```



The histograms allow the observation of surface-level patterns in linguistic styles.

1. Sentence Quantity: The top three MBTI types with the highest average sentence counts are ESTJ (150.1), ENTP (147.2), and ESFP (147.1). In contrast, the bottom three are INFP (135.3), INFJ (136.1), and ISFJ (139.9). This may indicate a correlation between the Extroversion/Introversion type and the extent of verbal expression. However, the overall variation in sentence quantity is relatively minor, suggesting limited explanatory significance.
2. Word Count:

3. Upper-Case Ratio:

4. Reading Ease:

6.2.3 Sentiment Analysis

Sentiment analysis, or opinion mining, is an active area of study in the field of natural language processing that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text. There are various approaches to sentimental analysis. In this assignment, we choose VADER (Valence Aware Dictionary for sEntiment Reasoning) as our tool to analyze the difference of different MBTI personalities in emotional expression due to its superiority in social media context (Kiritchenko et al., 2014).

VADER uses a pre-constructed system of "lexicon + weighted rules" to score each word in the input text and generate an overall sentiment score. For every given text, VADER first separates the text into words. Then, it rates the words from -4 to 4 (4 as the most positive, and -4 as the most negative) according to its 7,500-entity valence lexicon. To increase the rating accuracy and distinguish between strength of emotions, VADER corrects the scores with a set of rules, which includes increasing or decreasing the score when identifying capitalization, negative and contrastive words. The details rating process is represented in the flow chart below.

6.2.4 Exploratory Data Analysis (EDA)

```
1 def get_sentence_quantity(self):
2     for post in self.data["posts"].values:
3         self.basic_identities["sentence_quantity"].append(len(post))
4     self.basic_identities["ave_sentence_quantity"] = ave(self.basic_identities["sentence_quantity"])
```

References

- Boyle, G. J. (1995). Myers-briggs type indicator (mbti): Some psychometric limitations. *Humanities Social Sciences papers*, 30. <https://doi.org/10.1111/j.1742-9544.1995.tb01750.x>
- Fleenor, J. W. (1997). The relationship between the mbti and measures of personality and performance in management groups. *Developing leaders: Research and applications in psychological type and leadership development*, 115–138.

- Kiritchenko, S., Zhu, X., & Mohammad, S. (2014). Sentiment analysis of short informal text. *The Journal of Artificial Intelligence Research (JAIR)*, 50. <https://doi.org/10.1613/jair.4272>
- Yang, Y. (2022). Research on the application of mbti in organization. *2022 7th International Conference on Social Sciences and Economic Development (ICSSSED 2022)*, 1751–1754.
- Zibran, M. F. (2007). Chi-squared test of independence. *Department of Computer Science, University of Calgary, Alberta, Canada, 1(1)*, 1–7.