Word Count without reference	1858
Word Count including reference and appendix	2453
Exceeded word	0
Penalty	0

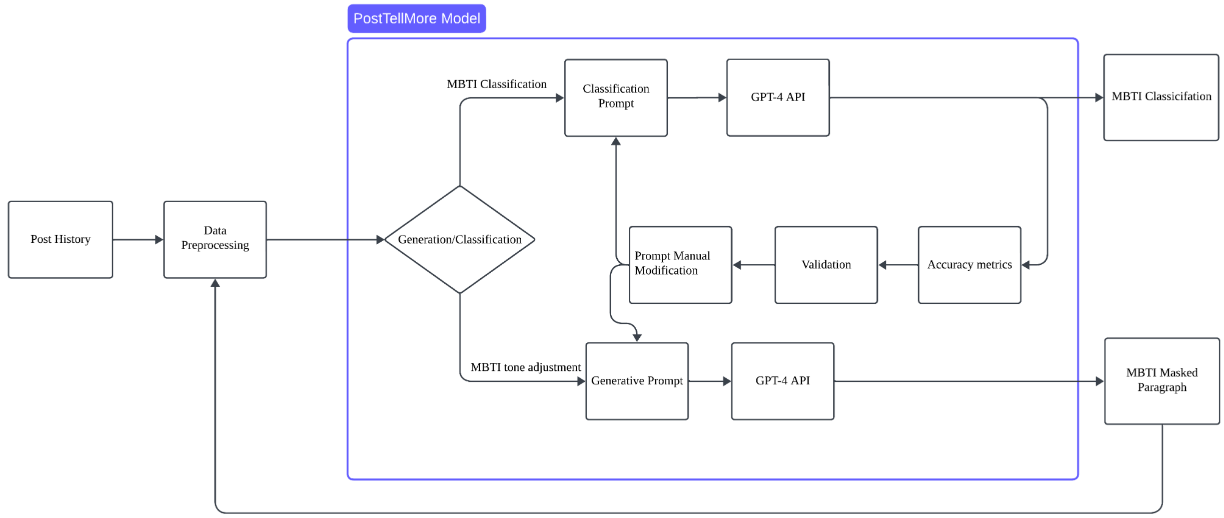
PostTellMore Final Report

Eric Liu & Haochen Zheng

Introduction

Myers-Briggs Type Indicator (MBTI) is based on Carl Jung's theory of psychological type. It indicates an individual's personality preferences in four perspectives: Extraversion (E) or Introversion (I), Sensing (S) or Intuition (N), Thinking (T) or Feeling (F) and Judging (J) or Perceiving (P). With the boom in the media, people seek to share their experiences, thoughts, and emotions on the internet, a social phenomenon that serves as a reflection of their personalities. However, this practice raises substantial concerns regarding privacy and security, given that personal traits are inadvertently disclosed to the public through these online posts. Consequently, our project centers on the meticulous evaluation of individuals' personalities and the creation of masked posts. The masked posts deliver similar content but in a different MBTI personality tone, addressing the critical need for reliable personality analysis and enhancing privacy security from online media interactions. The post history from a single person would contain a large number of words and symbols, making it hard to extract the personality signatures from the posts and most of them would not relate to the MBTI questionnaire. Therefore, machine learning and large language models would be introduced to capture potential signatures from massive paragraphs.

Illustration



Background & Related Work

A set of methods, Naive Bayes, Support Vector Machine, Decision Tree, Multilayer Perceptron, K-Nearest Neighbours and random forest has been utilized to determine the MBTI type based on the twitter data [1]. The Twitter data includes two categories. The first one is the behavioural category consisting of several tweets, followers, followed, favourites, listings and being favourited, the second one is the grammar category, which considers information from oNLP, MRC, sTaggers and LIWC. [2] The result of this study is that the RandomForest method achieves about 80% accuracy in MBTI personality under LIWC and oNLP. This result suggests that it is possible to determine the users' personalities based on what they write in social media. The posts shall contain enough information to determine the users' MBTI types. This shows the potential abilities of LLM to perform personality analysis based on users' posts.

Data and Data Processing

The dataset used in this project mainly comes from the PersonalityCafe forum[3], as it provides a large selection of people and their MBTI personality type, as well as what they have written. There are more than 8000 rows and each row contains about 50 posts from a user. There are about 1600 - 1800 words per row.

Below are the steps we follow:

1. Clean up http references from each post, and replace them with "url" to represent information that users quote web addresses in their posts.
2. There were tags in posts like #WhatAWonderfulDay. Which would confuse the tokenized process.

    a. Remove tag mark #.
    b. Separate the tag word if possible
    c. Modified it to become ‘What A Wonderful Day’ for example ‘#WhatAWonderfulDay’
3. Remove symbols like $#@& that either mislead phrases or are completely non-sense.
4. Delete all extra space, making the sentence clean.
5. Encode The MBTI type into four binary variables for further classification
6. Split data into Train and Test files with a ratio of 0.2

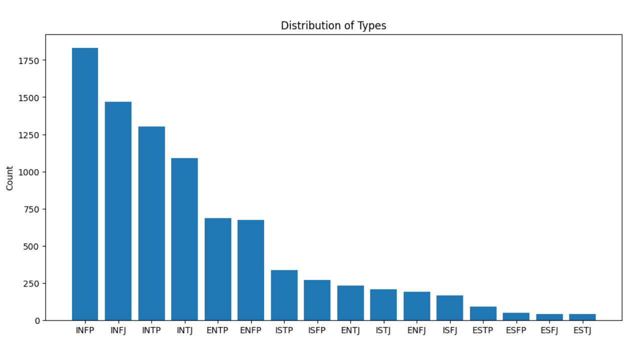
Here is a simple example of context.
Before:

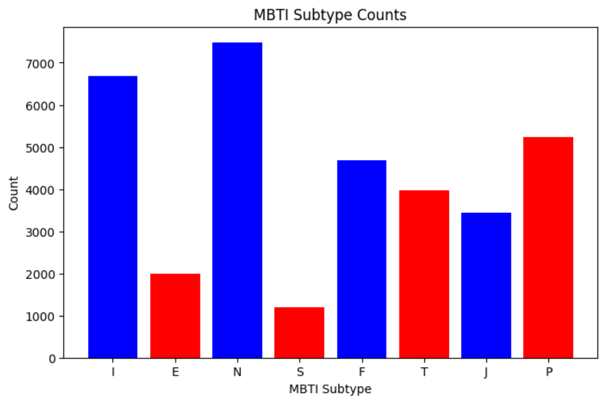
```
I started to make comics about turtle Gordon and unicorn Chimes -
here you can see two first stories: https://www.tumblr.com/blog/-alexxxandra |||
i fucking HATE when you INFPs ignore me, I'm all like
idgieyfviuefvifegedfiedgfipwgfwpifgepifgdig3294629472974274529!!@#@#@)#(@)(#&@
and you'll go *blink* *blink* and i'm like WHAT THE... |||
Cool ! Thank you ! yes , send it to turabo40@gmail.com ||
#WhyNotCapitalizeYourWordsToMakeItEasierToRead #HashBrownSelfie|||
```

After:

```
i started to make comics about turtle gordon and unicorn chimes
here you can see two first stories url
i fucking hate when you infps ignore me i m all like
and you ll go blink blink and i m like what the
cool thank you yes send it to email
why not capitalize your words to make it easier to read hash brown selfie
```

After these steps, sentence lengths are shorter by 20% and all words inside can be easily tokenized.





The dataset is unbalanced as the graph shows above. The smallest data group is ESTJ containing 39 samples.

To reduce bias, undersampling and oversampling are used.

- Undersampling: Reduce the size of each MBTI type to force the size of each type to match with each other.
- Oversampling: Split same-type posts into lines and shuffle to generate new phrases. It could generate some new samples to increase the size of the data.

The target value (MBTI type) would be split into four categories to make classification metrics more convenient.

Here is the table of preprocessed data:

type	posts	clean_text	E/I	N/S	T/F	J/P	mbti_label
INFJ	'http://ww	url and mc	FALSE	TRUE	FALSE	TRUE	[0, 1, 0, 1]
ENTP	'I'm finding	i m finding	TRUE	TRUE	TRUE	FALSE	[1, 1, 1, 0]
INTP	'Good one	good one	FALSE	TRUE	TRUE	FALSE	[0, 1, 1, 0]
INTJ	'Dear INTP	dear i enjc	FALSE	TRUE	TRUE	TRUE	[0, 1, 1, 1]
ENTJ	'You're fire	you re fire	TRUE	TRUE	TRUE	TRUE	[1, 1, 1, 1]
INTJ	'18/37 @.(	18 37 scie	FALSE	TRUE	TRUE	TRUE	[0, 1, 1, 1]
INFJ	'No, I can't	no i can t	FALSE	TRUE	FALSE	TRUE	[0, 1, 0, 1]
INTJ	'I tend to b	i tend to b	FALSE	TRUE	TRUE	TRUE	[0, 1, 1, 1]
INFJ	I'm not sur	i m not sur	FALSE	TRUE	FALSE	TRUE	[0, 1, 0, 1]
INTP	'https://wv	url i m in t	FALSE	TRUE	TRUE	FALSE	[0, 1, 1, 0]
INFJ	'One time	one time i	FALSE	TRUE	FALSE	TRUE	[0, 1, 0, 1]
ENFJ	'https://wv	url 51 o i v	TRUE	TRUE	FALSE	TRUE	[1, 1, 0, 1]
INFJ	'Joe santa	joe santag	FALSE	TRUE	FALSE	TRUE	[0, 1, 0, 1]
INTJ	'Fair enoug	fair enoug	FALSE	TRUE	TRUE	TRUE	[0, 1, 1, 1]
INTP	'Basically t	basically t	FALSE	TRUE	TRUE	FALSE	[0, 1, 1, 0]
INTP	'Your comi	your comi	FALSE	TRUE	TRUE	FALSE	[0, 1, 1, 0]

Architecture and Software

The raw data from the MBTI dataset is collected and passed to the data processor. After that, the processed data is allocated to our model.

Based on user choice, the model would decide if to do the classification only or generate a new post with a different tone and classify the personality in the new post. To perform classification only, use the classification prompt (see Appendix A). Referring to the designed prompt (see Appendix B), users can choose a preferred altered MBTI type by replacing the XXXX. The output contains two parts, the first part is MBTI types of original inputs and the second part is the altered posts that contain similar information but different MBTI tones. They are passed to different evaluation metrics to verify the results. Classification would pass the accuracy metrics directly, while generated posts would be classified to make sure the personality changed and go through manual checks to confirm the content/meaning is stable.

A judgement of qualifications will be applied. It includes 2 parts, whether generated posts contain similar information as the original one, this is a hard requirement, and whether the generated response has a tone of a different MBTI personality, the pre-trained classification prompt will be applied to label the newly generated texts and the number of different categories will be calculated to visualize the success.

Baseline Model or Comparison

The Baseline Model is based on the word embedding of GPT-2, performing classification with four labels according to the four categories of MBTI. A custom wrapper class is added to add the sigmoid activation function to it. Since the classification of each category shall be zero and one. It uses a fine-tuning method from HuggingFace to perform a prediction on the MBTI personalities. The output will be a clear prediction of the MBTI personalities.

The comparison method experienced several changes. We separate evaluation into 3 different categories: general accuracy, which represents all predictions, and categorical scores, which represent the predictions of four categories of MBTI individually. Since the test data is balanced with an equal number of opposite types and the accuracy is far above 50%, f1-score is given up.

Quantitative Results

The MBTI targets have four categories, meaning the traditional binary metrics in classification are not applicable. New metrics of accuracy are introduced based on four dimensions of categories. The general accuracy would be the mean accuracy of four separate categories.

Example: If the target is INFJ and the output is ENFJ, the general accuracy of that sample would be 0.75.

Classification	Baseline Model	Main Model
General Accuracy	71.04%	88.91%
Complete Correct Prediction[1]	27.04%	71.63%
Prediction with at most One Error	67.02%	91.25%

[1] Complete Correct Perdiction stand for the case four categories are all correct in prediction/

Extraversion (E) / Introversion (I)	72.87%	88.75%
Sensing (S) / Intuition (N)	70.77%	81.88%
Thinking (T) / Feeling (F)	73.94	92.50%
Judging (J) / Perceiving (P)	66.55%	92.50%

The table above shows the accuracy of classification prompting with both baseline and main models. The baseline model used 1600 samples took 5 epochs to train, and converged at about 71% accuracy. The main model was tested with 160 samples and the overall correctness is 17% higher. The main model took advantage of all regions compared with the baseline. According to the new metrics, the general accuracy equals the average classification accuracy of the four categories alone. Over 70% of outputs in the baseline model were not completely correct. Even though single-category classifications have acceptable accuracy in the baseline model, it was almost impossible to use that result as a reference. In terms of the main model, Over 70% of MBTI types are predicted completely accurately, proving that our prompt was able to capture information and guide GPT-4 to respect MBTI personalities.

When it comes to tone modification, the newly generated posts would be used as input in the classification part of the main mode. Based on the high accuracy in main model prediction, the classification part is capable of distinguishing if the generated tone differs from the original MBTI types. Since the content of the post can not be modified, some of the personality types can not be changed easily by adjusting the way of speaking, the ratio of how different the personality it could generate is listed in the table below. The highest possible personality it would be able to generate is two categories that differ from the original MBTI type.

N categories	Ratio of difference
0	0%
1	18.75%
2	56.25%
3	37.50%

4	6.25%

Qualitative Results（E）

Here is a success example:
Original Input:

```
you guys are awesome truly appreciate all your responses
it s been helping me accept this lifestyle because at
the end of the day i wouldn t change who i am for even
a second happy
```

Label: ENFJ
Generated Texts:

```
I must express my profound gratitude for your
presence in my life. Your responses have
brought me immense joy and comfort. I cannot
help but feel a sense of peace and tranquility
when I engage with all of you.
```

Generated Label: INFP

The hard requirement is whether the context is similar. In this example, both texts explain the thankfulness to friends due to their help to the user who is trying to accommodate his/her current life. The original text is clear i type using “you” while the generated text is more e type using “i”. The first one focuses more on an individual's idea, which is a “J” sign, while the second one focuses on another’s contribution which is a “P” type.

Here is a relatively worse result:

Original Input:

```
i was pretty straightforward about it i would
tell her that she needs to throw out old food
because it gets stale or it rots or i ll tell
her she needs to show her trash away her messes
```

Label: ESTJ
Generated Output:

```
I believe in being straightforward and direct
when it comes to matters like throwing out old
food or dealing with messes. It's important to
recognize that stale or rotten food can cause
problems, so addressing it is necessary.
Similarly, keeping our surroundings clean by
disposing of trash is essential.
```

Label: ESTJ

The generated text does not change to a new MBTI version. The potential reason could be that we are requiring the GPT-4 to keep the key information, which makes it harder to alter while the same information is delivered.

Discussion and Learnings

According to the qualitative and quantitative results, the classification performance of the main model is beyond our expectations, it was able to not only capture each category, but there are some potential connections between them that it can understand, making it better performance in complete correct prediction. As for tone modification, it was under our expectations, the personality would also reflect on the content of the posting, making it more difficult to switch the tone. Besides, the tone generation would take way more time than expected, forcing us to reduce the length of posts.

If we were able to start a next generation or a similar project, there are several points we should improve. We should collect data by ourselves to make the dataset more balanced, instead of doing oversampling which may still have bias or distract the prediction. Posts with more information should weigh more and reduce the total number of posts per sample, 50 posts per sample cost too much when it comes to a larger dataset. We would try to find a better way to preprocess the data including url, email, and symbols, some of them can still contain information about the person. An UI needs to be implemented so that we can demo the project making it more convincing.

Individual Contributions

We evenly distributed tasks and helped each other according to the workload and trouble we faced.

Accomplishment	Contributor
Search appropriate source of data	Eric Liu, Haochen Zheng
Proposal report	Eric Liu, Haochen Zheng
Preprocess and clean up data	Haochen Zheng
Baseline model implementation	Eric Liu
Baseline model fine-tune and testing	Eric Liu, Haochen Zheng
Main model implementation	Haochen Zheng
Progress report	Eric Liu, Haochen Zheng
Progress presentation	Eric Liu, Haochen Zheng
Fine-tune classification prompt	Haochen Zheng
Validation metrics	Eric Liu, Haochen Zheng
Fine-tune generation prompt	Eric Liu
Test and verify tone generation	Eric Liu, Haochen Zheng
Presentation slides and report	Eric Liu, Haochen Zheng

	Haochen Zheng	Eric Liu
permission to post video	Yes	Yes
permission to post final report	Yes	Yes
permission to post source code	No	No

Appendix A

Given the client's post history, predict their MBTI personality type. The website address is the post has been replaced by url, the upper cases at the start of sentence are converted to the lower cases

Provide the most likely result with the probability of E/I, N/S, F/T, and J/P categories on a scale of 100, with explanation

Consider factors such as the client's communication style, preferences for information, and recurrent topics to make an informed assessment of their potential MBTI personality type.

Here are some hints:

Extraversion tends to be more outgoing and energized by external stimuli, they need more communication with others, even if it makes them exhausted ;

while those with a preference for Introversion are often more reserved and energized by internal reflection, they tend to stay alone, but it does not deny they engage in communication.

Sensing would pay more attention to discussing things they observed or heard from others;

while Intuition or N means future-oriented and often thinking beyond the immediate reality.

Feeling means they are easier to feel empathy for others' emotions;

while Thinking personality may prioritize rationality over personal considerations.

Judging prefers to plan ahead, make decisions and reach conclusions.

while Perceiving is comfortable leaving decisions open and exploring possibilities.

Appendix B

Given the client's post history, predict their MBTI personality type.

The website address in the post has been replaced by url, the upper cases at the start of a sentence are converted to the lower cases

Provide the most likely result with the probability of E/I, N/S, F/T, and J/P categories on a scale of 100, with explanation

Consider factors such as the client's communication style, preferences for information, and recurrent topics to make an informed assessment of their potential MBTI personality type.

After the analysis of MBTI type, regenerate the given text in the tone of the new MBTI type: XXXX by changing the tone, style, perspectives and focuses, and keeping the key information delivered by the posts.

Here are some hints:

Extraversion tends to be more outgoing and energized by external stimuli, they need more communication with others, even if it makes them exhausted;

while those with a preference for Introversion are often more reserved and energized by internal reflection, they tend to stay alone, but it does not deny they engage in communication.

Sensing would pay more attention to discussing things they observed or heard from others;

while Intuition or N means future-oriented and often thinking beyond the immediate reality.

Feeling means they are easier to feel empathy for others' emotions;

while Thinking personality may prioritize rationality over personal considerations.

Judging prefers to plan ahead, make decisions and reach conclusions.

while Perceiving is comfortable leaving decisions open and exploring possibilities.

Reference:

[1]Lima, A. C. E. S., & De Castro, L. N. (2019). Tecla: A temperament and psychological type prediction framework from Twitter data. PloS One, 14(3), e0212844–e0212844. https://doi.org/10.1371/journal.pone.0212844
[2]Lima, A. C. E. S., de Castro, L. N., & Corchado, J. M. (2015). A polarity analysis framework for Twitter messages. Applied Mathematics and Computation, 270, 756–767. https://doi.org/10.1016/j.amc.2015.08.059
[3](MBTI) Myers-Briggs Personality Type Dataset https://www.kaggle.com/datasets/datasnaek/mbti-type