MBTI Personalities Prediction based on Fine Tuning BERT with Six Semantic Dimensional Database

######## and ########
Department of Computer Science and Electrical Computer Engineering
University of Michigan
Ann Arbor, MI, USA

1 Introduction

The Myers-Briggs Type Indicator (MBTI) is an instrument that measures and describes an individual's psychological type and is designed to help people understand their preferences, behavioral patterns, and ways of interacting with others. Currently, the official MBTI personality measure is a standardized questionnaire, which makes the test consistent across individuals. However, standardized questionnaires have several obvious drawbacks: 1) Based on the standardization of the questions, when conducting multiple tests, the questions can be predicted, which can affect the results of the questionnaire 2) Questionnaires usually have 3 to 5 options, making the degree indicated by the options more discrete 3) The tester's perception of the test questions leads to a bias in the results. Therefore, the purpose of this project is to predict MTBI personality through linguistic modeling by providing a response window for the test taker in the form of a self-description.

This approach is intended to address several of the aforementioned shortcomings: 1) the tester cannot predict what portion of the text will affect the personality test when self-describing 2) the representation of text can be a more continuous expression 3) there are no differences in the tester's perceptions of the questions through self-description.

Completing the refinement of the MBTI test will provide a more versatile aid to the field of counseling. It will provide more information to individuals seeking psychological testing as well as to counselors. This project is intended to be used in the field of psychological research and testing and is intended to help improve psychological research aids.

MBTI personality plays an increasing role in daily life, work collaboration and other aspects, but currently, **researchers has not done enough in the people's MBTI identification of self-reported content**, and **the machine learning algorithms applied were still primary, resulting in the entire dataset performance did not reach a high point**. This project fine-tuned a **sophisticated BERT structure to predict the MBTI identification with a six semantic dimensional dataset, which was unprecedented**, And this project was a milestone for more in-depth MBTI research in the future.

This project utilized the MBTI text dataset as the raw data Choong (2021) for pre-processing the MBTI dataset, which was designed in a way that was expected to include, but not limited to, tokenizing, stemming, and selecting of the text. Meanwhile, the inspiration of the six-dimensional semantic database based on the SSDD datasetWang (2023) , which included the calculation of the weights of the psycho-cognitive aspects, was used to fine-tuning the output of pre-trained BERT model.

In terms of model training and language processing, this project utilized Pytorch training libraries to build and configure various models for training, outputting relevant personality labels, and training strategies including, but not limited to, the use of splitting databases for the training and test sets and the use of The training strategy includes, but was not limited to, cross validation strategy to emphasize not overfitting the training text. Finally, the score was calculated for both real and predicted data.

2 Related Work

Choong and Varathan (2021) used basic NLP feature extraction method and five of the basic machine learning classifiers, namely: Naïve Bayes, Nearest Neighbour, Support Vector Machine, Logistic Regression, and Random Forest based on MBTI dataset Choong (2021) to complete classification, While our work is more advanced applying

Data	Label
I was going to ...	ENTJ
The movie was bad...	INTJ
Hahaha ASAP PLZ...	ISTP
I am killing you ...	ESFJ
Sorry for that but...	INTP

Table 1: Example of MBTI dataset Choong (2021)

Bert model to classify the personalities.

Wang (2023) rated 17,940 commonly used Chinese words on six major semantic dimensions to represent semantic information. however this article was not involved in considering MBTI personality. If we add this sentiment enhancement strategy to the classification part, our final output will be theoretically better.

N. Cerkez and Skansi (2021) applied long short-term memory (LSTM) and convolutional neural network (CNN) in NLP improving the current results of the MBTI multi-class classification. This project builds on this article and aims to enhance the model preprocessing aspects to achieve better theoretical results.

Nisha et al. (2022) analyzed the personality posted on Twitter by sentimental analysis and applied traditional machine learning methods like Naive Bayes (NB), Support Vector Machine (SVM), and XGBoost classifier, achieving 78%, 80% and 85% respectively.

Ryan et al. (2023) applied synthetic minority oversampling technique (SMOTE) method integrated with Word2Vec method to balance the MBTI dataset, improved selected machine learning model's performance and finally achieved at most 0.8337 F1 score.

Sirasapalli and Malla (2023) found that Personality is playing an essential role in recommendation system, decision making and so on, and measured personality based on available text data on social media. A new data source mapping method was applied and combined with data fusion techniques, leading to accuracy of 87.89% and 0.924 F1 score.

3 Data

3.1 Dataset

Table 1 shows five examples of MBTI dataset Choong (2021). This dataset contains almost 8k data of MBTI dataset.

Table 2 shows five examples of SSDD dataset Wang (2023). This dataset contains 1,427,992 Chinese and 1,515,633 English words' comments and every word contains 6 dimensions or 7 features.The SSDD dataset conducted six labeling experiments on 17,980 words, each focusing on one semantic dimension (visual, motor, social, emotional, temporal, spatial). Each annotation experiment was divided into 18 sessions, each containing 1000 words (the last session had 940 words). The emotional dimension was labeled using a 13-point scale (-6 for very negative, 0 for neutral, 6 for very positive) and the remaining 5 dimensions (visual, motor, social, temporal, spatial) were labeled using a 7-point scale (1 for very low, 7 for very high).

This project used MBTI Dataset as our training data, where each row represented one data, and the first column represents text, and the second column represents label. the text was tokenized first, and then the dictionary we needed was composed according to stemming and selecting strategies(more steps will be illustrated in the methodology part). Finally in the fine-tuning part, the six semantic dimensional dataset was applied to weight the outputs of the pre-trained BERT model, so that our fine-tuning model can classify 16 MBTI personalities in six dimensions.

3.2 Data Pre-processing

3.2.1 SSDD and MBTI dataset

To adapt the MBTI dataset for analysis, this study employs a preprocessing routine that utilizes the English-Bert Dataset from the SSDD collections, which categorizes English words into six dimensions. This dataset comprises 930,668 words and serves as the vocabulary foundation for tokenizing the MBTI dataset content. An initial consideration is the treatment of words in the MBTI dataset that lack corresponding entries in the SSDD dataset as 'None' during network training. Upon tokenizing the entire MBTI content, we identified a unique set of roughly 60,000 tokens, which we then designated as the new vocabulary. This approach significantly enhances training efficiency, as the original vocabulary size was found to be unwieldy. Subsequently, this refined vocabulary is employed to re-tokenize content from the MBTI dataset. Regarding classification labels, the MBTI identifies four dichotomies: Introversion (I) versus Extraversion (E), Intuition (N) versus Sensing (S), Thinking (T) versus Feeling (F), and Perceiving (P) versus Judging (J), culminating in 16 distinct personality

word	Vision	Motor	Socialness	Emotion	Emotion_abs+1	Time	Space
take	4.133333	3.666667	1.500000	0.666667	1.666667	1.800000	3.000000
walk	3.200000	5.366667	1.200000	0.233333	1.233333	2.066667	4.966667
room	4.400000	2.400000	1.466667	0.233333	1.233333	1.066667	6.633333
eyedrop	5.233333	5.233333	1.266667	-3.333333	4.333333	1.233333	1.533333
relative	3.433333	1.966667	6.100000	1.266667	2.266667	1.100000	1.266667

Table 2: Example of SSDD dataset Wang (2023)

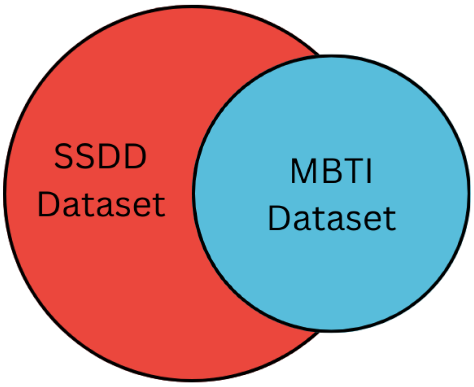

Figure 1: Relationship between SSDD dataset and MBTI dataset

types. However, each dichotomy operates independently, allowing for their representation as a four-dimensional one-hot encoded vector for loss calculation purposes. The dataset is partitioned into training and evaluation sets at an 80 % to 20% ratio, respectively, to facilitate model training and performance assessment.

Figure 1 illustrates the occurrence of certain words within the MBTI dataset that do not have counterparts in the English-Bert dataset. These words are subsequently categorized as 'Unknown' in the revised vocabulary text, reflecting their negligible contribution to network predictions. Furthermore, a notable consideration arises during the evaluation of MBTI test data: should a word, denoted as A, be present in the original English-Bert dataset but absent in the updated vocabulary text, it ostensibly represents valuable information. However, since the network's parameters pertinent to word A remain untrained throughout the learning process, its impact on testing performance is minimal, thereby justifying its exclusion.

3.2.2 Cleaning data

Some of the data for the MTBI prediction task contain irrelevant and extraneous information. In this project, links and emojis are not considered reliable indicators or useful features for inclusion in training because they do not represent any meaningful trends. Additionally, this project focuses on processing English text using the English-BERT SSDD dataset, which only includes English characters. Symbols and characters from other languages will be discarded to maintain a training process focused on data with six-dimensional significance, thereby helping the model converge more efficiently.

3.2.3 BERT Tokenizer

After data cleansing, this project utilizes the BERT tokenizer to adapt the text for the BERT model. The tokenizer breaks down sentences into segments that the pretrained BERT can recognize, reducing the occurrence of unknown words. The maximum token length is set to 512 to accommodate subsequent processing in the pipeline. Additionally, special tokens are added during truncation because the pretrained BERT model is designed to read sequences that start with '[CLS]' and end with '[SEP]'. For testing longer texts, the BERT Tokenizer will generate multiple fixed-length lists of tokens, each 512 tokens long, to be used in subsequent tasks.

3.2.4 Distribution and Resampling

The distribution of the data at the beginning was imbalanced(Figure 2). It is clear that for INFP, INFJ and INTP, their total share is more than half of the data, while for ISFJ and other types, their respective share only counts up to 1%. So a combination of upsampling and downsampling was chose to resample the original data, and put the number of original data into a customized interval, where upsampling means increasing the number of very few species by copying them into a set interval, and downsampling means randomly deleting a large number of species, thus controlling the number into the interval as well. The interval setting should satisfy the characteristic of the trend of the original data.

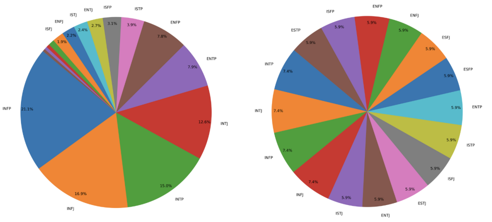

Figure 2: the Original Distribution (Left) and Resampled Distribution (Right)

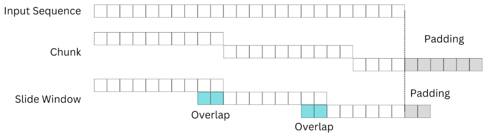

Figure 3: Chunk and Slide Window after tokenizer

3.2.5 Chunk and Slide Window

According to the characteristics of BERT, the length of its input sequence needs to meet the maximum length of 512. However, the average length of our mbti data is about 1,000 words, so we chose the slide window in chunk to do the processing of data after being tokenized. The difference between the slide window and original chunk is that slide window incorporates an overlap mechanism, which means that every time a chunk sequence is performed, a certain length of the previous chunk will be intercepted at the same time. For details, please refer to (Figure 3).

4 Methodology

At the first step, the pre-trained BERT model('bert-large-uncased') was called from Hugging Face. After getting the pre-trained model, we continued fine-tuning the model from the output of the model (the word embedding).

The general fine-tuning BERT model refers to after getting the embeddings of the paragraph (the word vectors of each word), if the task is to categorize the text, the special tokens [cls] (the vectors of the first token) were picked up to fill into a fully-connected layer representing a linear transformation, followed by a softmax layer, ending up with a base version of fine tuning of BERT.

Our improvement on this is that after obtaining the embeddings, we do not just use the word vector of the first word, but all vectors, and then brought in our SSDD database and broadcast it into the vectors, obtaining a three-dimensional data (1-dimensional is the sequence length, and 2-dimensional is the BERT's own word embedding itself, and the 3-dimensions are the feature dimension of SSDD). Then a softmax is attached to transform the linear into nonlinear and the final result is taken as an output(Figure 5).

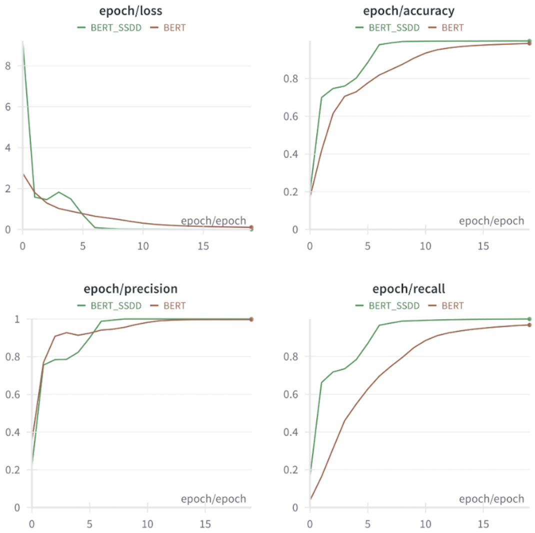

Figure 4: pipeline of fine-tuning BERT

Model	F1 score	Accuracy
BERT	0.695	0.695
BERT + SSDD	0.712	0.743

Table 3: metrics of BERT and BERT+SSDD

5 Results

5.1 Accuracy and F1-score

This project used two control methods, BERT and BERT+SSDD, to test whether SSDD can give BERT a boost. After the experiment, it is found that BERT can finally achieve an accuracy of 0.695, and an F1-score of 0.695, what is more, BERT+SSDD can achieve an accuracy and F1-score of 0.712 and 0.743(Table 5).

5.2 Training Process

According to the Figure 4, it is clear that BERT with SSDD converged faster than the original one, by about 5 epochs; the training accuracy, precision and recall indicators increased quickly before 5 epochs with small level of fluctuation, all achieving approximately 99% at the 6th epochs. However, the original BERT converged more stable, achieving around 98% eventually.

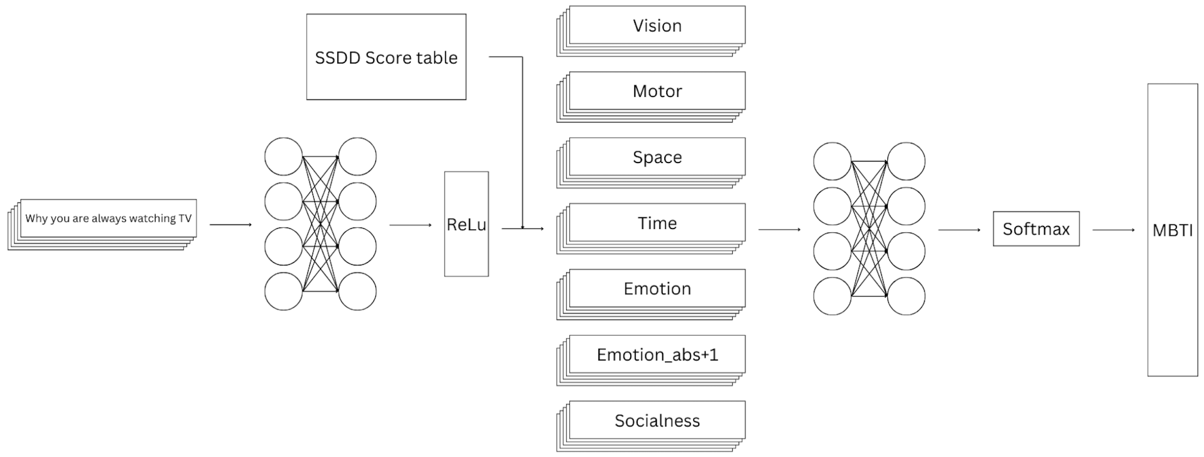

Figure 5: pipeline of fine-tuning BERT

6 Discussion

The application of the SSDD dataset significantly enhances both the F-1 score and accuracy. The pipeline is structured such that after focusing on contextual words, the output can effectively classify MBTI personality types. Adding a layer that processes the SSDD dataset and resembles word outputs enables the model to concentrate on six key features related to personality analysis.

However, there are several considerations regarding the dataset. First, data chunking involves splitting and segmenting long texts, which populates the training data effectively. However, this method introduces variability where some chunks may strongly suggest MBTI personality traits while others may only convey objective facts, thus acting as 'noise' in the training process and potentially degrading model performance. Another issue is data bias. In traditional MBTI analysis, the four dichotomies are continuous, meaning that personality traits are not strictly binary but are instead represented on a spectrum, such as 60% Extraversion (E) and 40% Introversion (I). However, the true labels is discrete.

Regarding model design, a significant challenge arises when words exist in the BERT vocabulary but are absent from the SSDD score table. The decision on how to score these words for their relevance in recognizing MBTI traits is contentious. In this project, words not observed in the SSDD dataset are assigned a zero score across all six dimensions, potentially overlooking words that might be strong indicators of personality traits.

7 Conclusion

Our research demonstrates that textual content, as opposed to traditional multiple-choice questionnaires, can serve as a valuable resource for analyzing MBTI personality types. The BERT and BERT + SSDD methodologies effectively uncover the implicit states within text and utilize these insights for personality prediction. Notably, the integration of more data indicators in conjunction with NLP tasks shows great promise in the field of psychology. The model's efficacy could be further enhanced by incorporating additional explainable indicators. This work could assist professionals by enabling rapid, real-time analysis and by using textual data to enrich their datasets.

8 Other Things We Tried

When considering the utilization of the SSDD dataset, our initial approach was to treat it as the primary vocabulary. However, this approach proved suboptimal, as treating the SSDD dataset as a closed vocabulary caused any new or unseen text to be categorized under the <UNK> (unknown) token, which detrimentally affected model performance.

Additionally, we experimented with a simple classifier as a baseline for MBTI prediction. Unfortunately, this classifier struggled with long sequences and tended to produce random predictions without convergence in loss. We also explored the use of clustering methods for self-labeling. However, it was challenging to correlate specific clusters with appropriate labels. Furthermore, we observed that cluster centers were often too close to each other, leading to instability in the classification process.

9 What You Would Have Done Differently or Next

Due to our lack of time, we only applied the SSDD dataset to fine-tuning part, but we found that if we could also add the SSDD dataset to the pre-train BERT, then theoretically it would be more effective.

If there is a next time, we would also like to try the truncating approach before a fully connected layer instead of chunking. Because this can ensure that there will be special tokens in each sentence, so that we can better utilize the special tokens' own embedding to do text analysis.

References

E. J. Choong. 2021. 8k mbti dataset from personality cafe. *figshare*.

E. J. Choong and K. D. Varathan. 2021. Predicting judging-perceiving of myers-briggs type indicator (mbti) in online social forum. *PeerJ 9*.

B. Vrdoljak N. Cerkez and S. Skansi. 2021. A method for mbti classification based on impact of class components. *IEEE Access, vol. 9*.

Kulsum Akter Nisha, Umme Kulsum, Saifur Rahman, Md Farhad Hossain, Partha Chakraborty, and Tanupriya Choudhury. 2022. A comparative analysis of machine learning approaches in personality prediction using mbti. In *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2021*, pages 13–23. Springer.

Gregorius Ryan, Pricillia Katarina, and Derwin Suhartono. 2023. Mbti personality prediction using machine learning and smote for balancing data based on statement sentences. *Information*, 14(4):217.

Joshua Johnson Sirasapalli and Ramakrishna Murty Malla. 2023. A deep learning approach to text-based personality prediction using multiple data sources mapping. *Neural Computing and Applications*, 35(28):20619–20630.

Zhang Y. Shi W.et al Wang, S. 2023. A large dataset of semantic ratings and its computational extension. *Sci Data 10*.