Story to MBTI

Group members: Malcolm Zhao, Weixiao Wang, Lingxin Li, Yun Tang, Binglan Lin, Yuan Lu

1 Background

Understanding characters' personalities is indispensable for grasping a story's greater message, engaging audiences, and analyzing films. Since characters shape the storyline through their conversations, actions, and decisions, they guide viewers through the intricacies of the plot and elicit emotional responses that deepen the viewing experience. Additionally, the Myers-Briggs Type Indicator (MBTI) is a popular personality scale in the field of psychology that assesses psychological preferences based on Carl Jung's theory. It aims to reveal how people perceive the world and make decisions in four dimensions:

- **E/I**: Extraverts (E) tend to focus on the external world and gain energy from it; Introverts (I) tend to focus on their internal world and derive energy from within.
- **S/N**: Sensing (S) focuses on concrete, practical details and real-world situations; Intuition (N) focuses on abstract concepts and possibilities.
- **T/F**: Thinking (T) prioritizes logic and objective principles when making decisions; Feeling (F) prioritizes subjective and interpersonal considerations.
- **J/P**: Judging (J) prefers a structured and planned lifestyle; Perception (P) type prefers flexibility and keeping options open.

The goal of *Stroy2MBTI* is to predict a movie character's MBTI personality types based on the narratives of the character.[1] By leveraging NLP models, we can uncover nuanced personality traits embedded within movie scripts. These insights into characters' personalities facilitate a deeper understanding of their intricacies, fostering stronger connections among filmmakers, viewers, and characters. It enhances character development by aligning motivations and actions with realistic traits, leading to deeper viewer engagement as audiences connect with familiar or contrasting personalities. Additionally, with heightened understanding, writers and actors are empowered to portray characters authentically, ensuring alignment with the script's themes and dynamics. Furthermore, by implementing such insights into characters' personalities, movie platforms can provide personalized recommendations. The personality information enhances content recommendations, enabling platforms to suggest films that resonate with viewer preferences.

2 Dataset and Preprocessing

The dataset we used in this study is provided by Sang et al.'s (2022) research on MBTI Personality Prediction for Fictional Characters Using Movie Scripts[2]. It comprises textual dialogues from 3,543 characters in 507 movies, along with the corresponding character names and movie names. The numerical MBTI scores represent the level of four dimensions as outlined in the preceding section. The structured presentation of movie scripts sets the dataset into two distinct parts: verbal dialogues and non-verbal scene descriptions, facilitating a comprehensive narrative analysis. For a comprehensive overview of features, explanations, and data types, please consult Table 1.

[1] Our code and data are released at https://github.com/MalcolmZhao/berkeley_ieor242B.git

[2] Sang, Y., Mou, X., Yu, M., Wang, D., Li, J., & Stanton, J. (2022). MBTI personality prediction for fictional characters using movie scripts. *arXiv preprint arXiv:2210.10994*.

Features	Explanation	Data Type
id	unique identification	int64
mbti_profile	character name	object
subcategory	movie name	object
vote_count_mbti	number of voters	int64
I/N/F/P/E/S/T/J	score of each channel	int64
dialog_text	actual words the character speak	object
scene_text	scene description of actions	object
mention_text	text mentioning the character	object

Table 1: Dataset Columns and Explanation

To clean noise from textual and numerical data and normalize them for use in the subsequent models, several preprocessing steps are applied:

- **Remove invalid characters:** Considering the necessity of utilizing dialogue to predict characters' personalities, it is crucial to ensure an adequate amount of text data for modeling. Thus, we remove characters with less than 80 words, those lacking unique dialogue, and those with null values in more than 3 channels of the personality.
- **Clean text**: Recognizing the presence of hashtags within the text data, such as "he ##llo", regular expressions are employed to remove these special characters, ensuring consistency and usability of text data.
- **Remove stopwords**: From a psychological point of view, stopwords have few meanings that indicate one's personality. As a result, NLTK's stopwords module is utilized to filter out stopwords and punctuations from text.
- **Fill in null value**: In each personality dimension, missing values are replaced with the median value calculated from the corresponding dimension of other characters.
- **Tokenizing and embedding:** In the following part, we embark on diverse models, each employing different methodologies such as word2vec embedding and pre-trained BERT models imported from Hugging Face to tokenize.

Following preprocessing steps and removal of unnecessary characters, we obtain a dataset consisting of 2,144 valid character entries, with each entry encompassing 9 distinct features. These features include basic character information, scores for four personality channels, and dialogue text, setting the stage for subsequent model construction.

3 Model Construction

In the following model construction process, we explore various deep learning models and architectural considerations tailored to our textual data. This involves meticulous training procedures, including data partitioning, loss function selection, experimentation with models, and fine-tuning of hyperparameters.

The entire dataset is partitioned into three segments: the training set, comprising 80% of the total dataset; the validation set, encompassing 10%, for hyperparameter selection, and the remaining 10% designated for testing purposes. Given the potential variability in dialogue length, we pivot our attention towards leveraging attention mechanisms for predicting MBTI from stories. These mechanisms enable the model to selectively attend to pertinent segments of the narrative while maintaining a holistic understanding of the entire context, facilitating the capture of nuanced relationships between story components. Thus, we consider both the Transformer model, which incorporates positional encoding, and BERT, which employs masked language modeling to capture bidirectional contextual information effectively. Detailed models are as follows:

1. Untrained transformer coupled with word2vec embeddings to predict scores, resulting in roughly 50,000 vocab size
2. Base-BERT model (12 layers) sourced from Hugging Face for score prediction with the last 6 layers unfrozen for backpropagation (Text input was marked for classification purposes to avoid sequence prediction encodings. As restricted by the pre-trained Bert model, 512 words that appeared most frequently are kept for training purposes)
3. Large-BERT model (24 layers) provided by Hugging Face for classification across four distinct channels with the last 4 layers unfrozen for backpropagation (Similar to model 2, we kept 512 vocab size and preprocessed text input to mark it for classification).
4. Large-BERT model (24 layers) from Hugging Face to predict scores with the last 4 layers unfrozen for backpropagation (Still, we kept 512 vocab size for prediction and transformed text input to clarify its classification purposes)

For the loss function, we chose mean squared error (MSE) when dealing with regression problems, predicting scores in our models. And for classification tasks, such as model 3, to predict personality types directly, we utilized cross-entropy loss. We trained all the models with Adam optimizer, and each in a total of 400 epochs. The first transformer model is trained with a 10^{-4} learning rate, and the three BERT-based models with a 10^{-3} learning rate. Dropout and batch normalization are also integrated to enhance the model's generalization and stability during training. All models were trained on a system equipped with 2 Intel Xeon Platinum 8352Y CPUs (32c/64t @ 2.4GHz), 4 NVIDIA RTX A6000 GPUs, and 256GB ECC DDR4-3200 RAM.

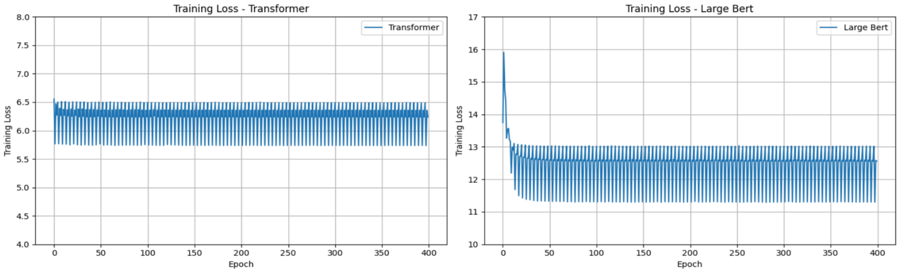

Figure 1: Training Loss for Transformer (model 1) and BERT models (model 4)

The two graphs above provide insights into the training dynamics of both the Transformer and BERT models. We need to point out that the batch size differs: 32 for the Transformer and 64 for BERT models, leading to different scales of training loss. It's foreseeable that the Transformer model struggles to converge within the designated training epochs due to data limitation. The performance of BERT models varies depending on their configurations and objectives. While the Base-BERT model with 12 layers and the Large-BERT model designed for classification encountered difficulties in achieving convergence and had poor performance on the validation set (less than 50% accuracy in most channels), indicating potential issues with the model architecture or training procedures, the Large-BERT models tailored for score prediction convergence within 30 epochs, and demonstrated more promising outcomes. This made sense as the dependent variable "score" provides more valuable information than simple classification results, and the pre-trained Large-Bert model is more capable of encoding the given input.

Since our models generate continuous numerical outputs, converting them into classifications requires establishing optimal benchmark thresholds for each personality trait channel. To achieve this, we rely on the F1 score metric computed using the validation dataset. By systematically exploring various threshold values from 0.1 to 0.9 with intervals of 0.01 and assessing their corresponding F1 scores, we can identify the thresholds that yield the optimal classifications for each personality trait, as illustrated in the subsequent table.

Channel	E/I	N/S	T/F	J/P
Threshold	0.23	0.1	0.33	0.22

Table 2: Threshold for Each Personality Channel

4 Synthesis of Results

We use macro-averaged F1 as our evaluation metric for comparison, which combines precision and recall, providing a balanced measure of the model's performance. Also, as Sang et al. put significant effort into evaluating their model's performance using F1, such selection naturally provides a baseline for this project (2022). Additionally, following the proposal's feedback, the accuracy of our BERT model is also offered to evaluate the proportion of correctly classified instances. The final result of the test set is presented in the following table 3.

A noteworthy observation from the results is that all four channels exhibit F1 scores surpassing 0.5, indicating a level of reliability in our model's predictions. Particularly striking is the performance of the T/F and J/P aspects, which achieve F1 scores exceeding 0.7, accompanied by an accuracy rate exceeding 55%. This suggests that our model excels in minimizing misclassifications and is very effective for a classifier, particularly for most applications. Furthermore, in comparison with the existing solutions proposed by Sang et al., as illustrated in Table 4, our model demonstrates superiority across all four prediction channels (2022). Although our model has yet to surpass human performance, it outperforms existing models in F1 scores, showcasing its potential for practical implementation and advancement in the field.

Metric	E/I	N/S	T/F	J/P
F1 score	61.33	58.39	75.53	74.47
Accuracy	45.33	36.23	59.55	56.75

Table 3: Final Result of Large-BERT Model (24 layers)

Model	E/I	N/S	T/F	J/P
SVM	54.65	55.41	52.83	56.18
BERT	56.06±0.73	55.59±3.36	57.13±0.97	57.59±1.40
MV-MR BERT	57.50±2.04	57.42±4.27	60.33±0.93	59.83±1.42
-multiview	57.30±1.91	57.05±1.80	57.04±2.05	57.32±2.21
Human Perf.	98.19±0.60	97.82±0.10	98.51±0.67	98.03±0.19

Table 4: Result Provided by Sang et al.

In conclusion, through meticulous data preprocessing steps and model construction, including attention mechanisms and transformer-based architectures, this project achieves promising results in predicting personality traits, surpassing existing models in F1 scores across all channels. It underscores our model's potential for practical implementation and advancement in the field of character analysis and film recommendation systems, beneficial for both viewers and film practitioners.

Building on our successful outcomes, there's room to explore the integration of other cutting-edge models to further enhance our predictions. For example, variants of BERT, such as RoBERTa and ALBERT, could provide insights into its effectiveness for personality prediction. GPT models are also renowned for their natural language understanding capabilities, which could capture intricate dialogue patterns and potentially improve prediction accuracy. Thus, GPT-prompted inputs may be a good choice for processing the first few layer's input and output. Moreover, our current project predominantly relies on dialogues for personality trait prediction, which might not capture the full spectrum of human behavior and personality. In further analysis, incorporating additional features beyond dialogues, such as character actions and interactions, could enhance the predictive power of the models.

References

Sang, Y., Mou, X., Yu, M., Wang, D., Li, J., & Stanton, J. (2022). MBTI personality prediction for fictional characters using movie scripts. *arXiv preprint arXiv:2210.10994*.