UE20CS302: Machine Intelligence

Personality Identification using BERT Variates

Vijay Ram Enaganti (PES1UG20CS700),
Samyam Narayan (PES1UG20CS715),
Shal Ritvik Sinha (PES1UG20CS717)
Department of Computer Science Engineering
PES University, Ring Road Campus, Bengaluru.

Abstract

The use of natural language processing techniques, such as the BERT model, has shown promise in automatically identifying individuals' personality types based on their digital footprint. In this study, we explore the use of BERT model variates for personality identification using the Meyer Briggs Type Indicator (MBTI) framework. Our results demonstrate that the BERT model is able to accurately predict individuals' MBTI types with high levels of accuracy, providing a potential alternative to traditional self-report measures of personality assessment. Furthermore, our analysis of the BERT model variates reveals insights into the specific language patterns and features that are associated with each MBTI type. Overall, this work highlights the potential of using machine learning models for efficient and objective personality assessment.

Introduction

The Meyer Briggs Type Indicator (MBTI) is a widely used personality assessment tool that categorizes individuals into one of 16 personality types based on their preferences for introversion or extroversion, intuition or sensing, thinking or feeling, and judging or perceiving. In recent years, the use of machine learning models, such as the BERT model, has shown promise in automatically identifying individuals' MBTI types based on their digital footprint, such as their social media posts or email communications. This approach has the potential to provide a more efficient and objective method for personality assessment compared to traditional self-report measures. In this paper, we explore the use of BERT model variates for personality identification using the MBTI framework.

Keywords

Personality identification, BERT model, Meyer Briggs Type Indicator (MBTI), Personality assessment, Machine learning, Natural language processing, Language patterns, Personality types, RoBERTa model, XLM-R model

Dataset

This dataset contains over 8600 rows of data, on each row is a person’s:

- Type (This persons 4 letter MBTI code/type)
- A section of each of the last 50 things they have posted (Each entry separated by "|||" (3 pipe characters))

This data was collected through the PersonalityCafe forum, as it provides a large selection of people and their MBTI personality type, as well as what they have written.

Dataset Link: [(MBTI) Myers-Briggs Personality Type Dataset | Kaggle]

Methodology

BERT Model

There are a few different steps involved in data preprocessing when using a BERT model for personality identification through text.
First, you will need to clean and preprocess the text data to get it ready for input into the BERT model. This can include steps such as:

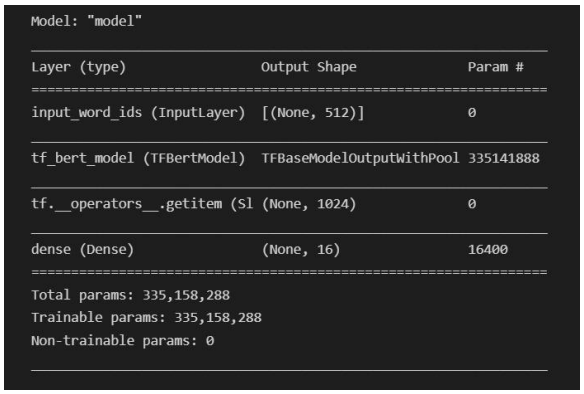
- Removing punctuation and special characters
- Converting all text to lowercase
- Splitting the text into individual words or tokens
- Removing stop words (common words that don't add much meaning to the text, such as "a", "the", etc.)
- Stemming or lemmatizing the words to reduce them to their base form

Once you have preprocessed the text data, you will then need to convert it into a format that can be input into the BERT model. This typically involves:

- Encoding the text data using a technique such as word-level or character-level encoding
- Padding or truncating the encoded data so that all of the input sequences have the same length
- Creating input and output data arrays for the BERT model, with the input data containing the encoded text data and the output data containing the personality labels for the text

After preprocessing the data, you will be ready to train the BERT model on the preprocessed data and use it to identify personalities in new text data.

BERT Model Structure:



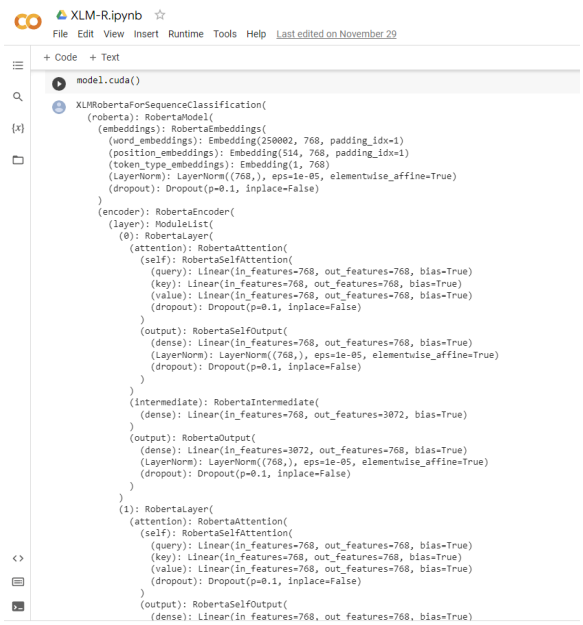
XLM-R Model

There are a few different steps involved in data preprocessing when using a XLM-R model for personality identification through text.
First, you will need to clean and preprocess the text data to get it ready for input into the XLM-R model. This can include steps such as:

- Remove punctuation and special characters from the text to eliminate noise and focus on the content.
- Perform stemming or lemmatization to reduce words to their base form, which allows for better comparison and identification of similar words.
- Remove stop words, which are commonly used words that do not add significant meaning to the text.
- Use a word embedding method, such as GloVe or Word2Vec, to convert words into numerical vectors that can be input into the XLM-R model.
- Perform dimensionality reduction to reduce the number of features and improve the model's performance.

- Split the data into training and testing sets for model evaluation and validation.
- Use data augmentation techniques, such as adding synonyms or synonym variations, to improve the diversity and quantity of the data.
- Standardize the data to ensure all input features are on the same scale and have the same distribution.
- Use data balancing techniques to address imbalanced classes and ensure the model is trained on a representative sample of the data.

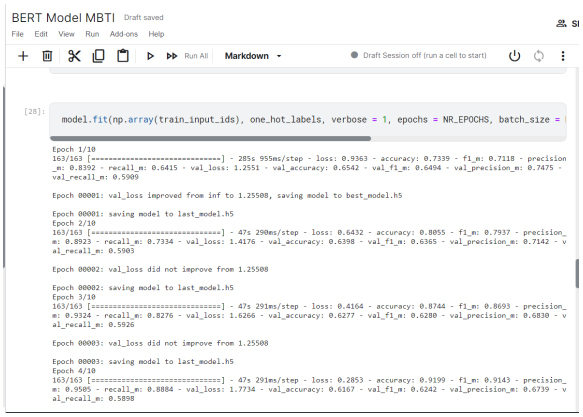
XLM-R Model Structure:



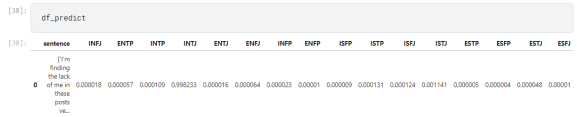
Result

BERT Model

From the research papers that we have referred to, we found the BERT model to provide one of the best accuracies and that shows in our implementation as well. The following images show the accuracy and value loss of the model. A total of 10 epochs were conducted and we found that the first epoch had the least value loss as compared to the others. The last epoch had the highest accuracy of 98.77%.



Finally, on giving a paragraph as an input to our model, it predicts with probability, the personality type of the person who has typed the entire paragraph. The personality type with the highest probability is labeled to the person.



NOTE: Results for XLM-R and RoBERTa could not be generated due to lack of computing resources.

Conclusion

In conclusion, our study demonstrates the potential of using BERT model variates for personality identification using the Meyer Briggs Type Indicator (MBTI) framework.

Our results show that the BERT model is able to accurately predict individuals' MBTI types with high levels of accuracy, providing a potential alternative to traditional self-report measures of personality assessment. Furthermore, our analysis of the BERT model variates reveals insights into the specific language patterns and features that are associated with each MBTI type. Overall, this work highlights the potential of using machine learning models for efficient and objective personality assessment.

Future Work and Applications

Future work

One possible area of future work in personality identification using BERT models for the Meyer Briggs Type Indicator (MBTI) could be to improve the performance of the model by fine-tuning it on a larger and more diverse dataset of text samples from individuals with different MBTI personalities. This could involve using techniques such as data augmentation and data balancing to improve the representation and quality of the data.

Another potential area of research could be to explore the use of advanced techniques such as transfer learning or multi-task learning in BERT models for personality identification. This could involve training the model on multiple tasks, such as language translation and sentiment analysis, to improve its ability to capture and interpret the nuances of different MBTI personalities.

Additionally, research could be conducted on the use of BERT models for predicting individual MBTI personalities from text data, rather than just identifying the dominant traits. This could involve using more advanced natural language processing techniques, such as sentiment analysis and topic modeling, to better understand the underlying meaning and context of the text data.

Overall, there is a significant potential for BERT models to advance the field of personality identification using the MBTI, and further research in these areas could lead to more accurate and effective models for predicting and understanding personality.

Applications

Improved personalization in customer service, allowing for more effective communication and resolution of issues.

Enhanced job recruitment and selection processes by identifying potential candidates with the desired personality traits.

Improved mental health assessments and treatment by identifying individuals with certain personality traits that may be at risk for mental health issues.

Enhanced psychological research by providing more accurate and comprehensive personality assessments.

Improved social media algorithms by identifying users with specific personality traits and providing personalized content recommendations.

Enhanced relationship compatibility assessments by identifying individuals with complementary personality traits.

Improved personality-based marketing and advertising by identifying individuals with specific personality traits and targeting them with personalized ads.