MBTI personality type prediction based on recent writings

Raehyun Kim, Tianhao Wu, Shiping Xu

1. Background Information

The Myers Briggs Type Indicator(MBTI) is a personality type system that divides every into 16 distinct personality types. It is one of the most popular personality tests in the world in terms of its use. Despite the lack of concrete scientific evidence, the MBTI test could provide useful information in a wide range of areas.

Our goal is to predict MBTI personality type based on each individual's recent writings on social media. Considering that the formal MBTI test is pricey and shows low test-retest reliability, it would be more convincing and accessible if we can determine each individual's personality system based on their post by developing a model. The dataset given by the Kaggle competition((MBTI) Myers-Briggs Personality Type Dataset) consists of MBTI personality type and last 50 posts of 8600 people. To extract features from the posts, we are planning to apply techniques from Natural Language Processing(NLP). The details will be provided at Section 3.

2. Methods

Our task is a multiclass classification problem that falls under supervised learning. We will use several methods to train the models, then compare the results. The methods under consideration are as follows.

   a. Classical Machine learning:
      i. QDA
      ii. Logistic Regression
      iii. SVM
      iv. Decision Tree Classifier
      v. Random Forest
      vi. Boosting
   b. Neural networks:
      i. BERT: Powerful transformer model for NLP tasks
      ii. LSTM: Works well when multiple tweets are considered.
      iii. CNN
   c. Ensemble:
      i. Better predictive performance compare to single model

We will train models to classify :

   a. The complete MBTI type. E.g. INFJ
   b. Individual axes separately. E.g. I,N,F,J. Since the E/I axis has dramatic differences, whereas an axis like T/F has subtle differences.

3. Core of the work

Since we are only given 50 posts of each person in the raw dataset, our first contribution in this project is to preprocess the data and extract features from these texts. Our second major contribution would be to apply several machine learning models on this dataset and compare models' performance with cross validation. For each model, we will use validation to tune hyperparameters and select the set of features with highest accuracy. Therefore, we would expect to be judged primarily on our exploration of different models, hyperparameters and features, and our programming in general.