Project Proposal

Haotong Wu & Ragnor Wu

Title: Personality Type Prediction using Natural Language Processing and Machine Learning

Introduction:

The Myers-Briggs Type Indicator (MBTI) is a widely used personality test categorizing individuals into 16 distinct personality types. Despite its popularity, the validity of the MBTI has been questioned, especially in predicting behaviour and language styles. In this project, we aim to explore the predictive power of the MBTI by using natural language processing and machine learning to predict the personality type based on the analyze the writing style of individuals.

Objectives

- Develop a machine learning model that can predict an individual's personality type based on their writing style.
- Analyze the predictive power of the MBTI and identify any patterns or correlations between personality types and language styles.
- Evaluate the performance of the machine learning model and select the most appropriate model.

Methodology

The proposed methodology for this project is as follows:

1. Data Collection: The data is from Kaggle: https://www.kaggle.com/datasets/datasnaek/mbti-type/code, and the dataset of over 8600 rows are obtained from the PersonalityCafe forum, consisting of each individual's 4-letter MBTI code/type and a section of the last 50 things they have posted.
2. Data Preprocessing: The text data will be cleaned and preprocessed using techniques such as tokenization, stemming, and stopword removal. The data will be convert to lowercase, and any URLs or numbers will be replaced with a special token to standardize the data. And the features will be represented as bag-of-words and bag-of-ngrams, which are representations that count the frequency of words and groups of words (ngrams) in each writing sample.
3. Model Development: A variety of machine learning algorithms, such as logistic regression, decision trees, linear Support Vector classifier and naive bayes will be trained on the preprocessed data to predict an individual's personality type.
4. Model Evaluation: The performance of the machine learning models will be evaluated using classification evaluation metrics, with precision, recall, and F1-score. The performance will be tested on a separate validation dataset to ensure that it generalizes well to new data.

5. Visualization: The results will be visualized using techniques such as heatmaps, scatter plots, and bar graphs to identify any patterns or correlations between personality types and language styles.

Plan

- 1: Data Collection and Preprocessing
- 2: Model Development and Evaluation
- 3: Correlation Analysis and Visualization
- 4: Write the Final Report

Conclusion and Application

The results of this project can provide insights into the validity of the MBTI in predicting behaviour and language styles, and may have practical applications in areas such as recruitment and team building. For example, companies may use this approach to analyze the language styles of potential employees and determine if their personality types align with the company culture and the requirements of the job.