Text-Based Prediction of MBTI Personality Type

Project Category: Natural Language

Daniel Kim
SUNet ID: ddk
Department of Computer Science
Stanford University
ddk@stanford.edu

Jenny Mai
SUNet ID: mcmai
Department of Computer Science
Stanford University
mcmai@stanford.edu

Emily Nguyen
SUNet ID: eqnguyen
Department of Computer Science
Stanford University
eqnguyen@stanford.edu

1 Key Information to include

- External collaborators or mentors (if you have any): None
- Sharing project with another class: No

2 Motivation

Although the concept of personality tests has been around for over 100 years, all of these personality tests follow the same structure: classify the user's personality based on their answers to a fixed set of questions. However, we would like to explore how personality types, more specifically the Myers-Briggs Type Indicator (MBTI), can be indirectly inferred from a user's text. While doing background research on this topic, we came across several projects that classified the user's personality based on their text answers to open-ended questions. Our project aims to predict the user's MBTI personality type based on their tweets. This application has the potential to uncover other personality types indirectly, through whatever people say on the internet (rather than just Twitter).

3 Methodology

We will be using the (MBTI) Myers-Briggs Personality Type Dataset from Kaggle[1], which includes over 8600 rows of Twitter users' last 50 Twitter posts and their respective MBTI labels. To preprocess the textual data, we will use the "bag of words" model with normalized tf-idf representation to feed into our classifiers. As a baseline classifier, we intend to use a multinomial Naive Bayes classifier, while also exploring other classifiers such as linear Support Vector Machines (more TBD).

4 Intended experiments

We will evaluate our classifiers using the test mean squared errors from k-fold cross validation as well as evaluation metrics such as accuracy (number of correct examples / total number of examples), precision, and recall. We will also compare our classifiers using these metrics to find out which classifier did the best on our dataset.

[1] https://www.kaggle.com/datasets/datasnaek/mbti-type