Myers Briggs personality predicting

Matthew Rogers *
School of Computing
Clemson University
mwr2@clemson.edu

Yue Zhang
School of Computing
Clemson University
yzhng@clemson.edu

Abstract

The Myers Briggs personality assessment puts a personality type to an individual. There are four components to an individual's personality. For each component, an individual takes on one out of two attributes. Since each of the four components has two attributes, then there are sixteen different personality types that an individual can be labeled as. For the first component there is introversion vs. extroversion, for the second there is intuition vs. sensing, for the third thinking vs feeling, and for the last component there is judgment vs perception. The objective of this project is to investigate the effectiveness of the perceptron to label each component of an individual's personality type.

Due to the nature of the Myers-Briggs type , we can break down the classification task with 16 classes in to 4 binary classification tasks. This is because a MBTI type is composed of 4 binary classes, where each binary class represents a dimension of personality of the MBTI personality model as theorized by the inventors. Therefore, instead of training a multi-class classifier, we instead train 4 different binary classifiers, such that each specializes in one of the dimensions of personality.

Our main dataset set is a public Kaggle data set containing of 6939 rows of data. Each row represents an anonymous individual. For each individual there is a personality type feature and multiple features containing the posts of that individual. This dataset will be altered to better fit the goals of the project.

The first action that needs to be taken is to "split" each component of the personality type into separate features. For example, the introversion vs extroversion component will be one feature and the intuition vs sensing component will be another feature, etc.

Second, since the data set comes from Twitter, where individuals communicate strictly via written text, some word removal is necessary, since we want every word in the data to be as meaningful as possible. we need to remove some very common words like "a", "the", "or". etc. also we need to remove types themselves(eg 'INTJ', 'INFP', etc), so to prevent the model from cheating by learning to recognize mentions of MBTI by name.

Third, we need to transform words into their root word (e.g. "singing", "sang", "sung" all become "sing"). This will allow us to make use of the fact the infected forms of the same word still carry one shared meaning.

As can be noticed from the previous action, there will be a large number of features. To reduce the number of features, the highest frequency words for an attribute will only be used. It has not been decided what number of words will be allowed yet. For example, say that the first component is being looked at first, then in the training set fifty of the highest frequency used words for each individual would be chosen and then combining all of these words will be the features. A perceptron

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Under review.

will be trained for each component of the personality type. Then, for the test set, the featured words will be counted for each individual. The perceptron will then make a decision on each component. The accuracy of this algorithm will then be assessed. In addition to using words as features to predict personality type for each of the four attributes, a separate dataset will be used where the letter frequency average for each individuals set of posts will be used as features (26 features). A perceptron will still be used to attempt to classify a personality type for each attribute. The effectiveness will then be assessed between these two sets of features (words vs letters).