Capstone Proposal - Myers Brigg’s Personality Prediction from Text Data

Danny Clifford

April 9, 2018

Proposal

Domain Background

The study of personality has become increasingly popular in the past century, giving rise to personality models like the Myers Briggs Personality Index and Big Five Personality traits. These models have been used to predict many things such as job satisfaction and job placement, among others. In the early 1940’s, Isabel Briggs Myers and her mother were interested in Carl Jung’s book called Psychologyical Type, and creating the MBTI for research in psychology and educational testing.

The theory of MBTI is there are 16 personality archetypes that are divided into 4 dimensions.

1. Extroversion vs. Introversion - measuring the degree the individual looks inside or outside the self in reaction to their environment.
2. Sensing vs. Intuition - measuring the degree of preference for external perception or intuitive feelings in determining truth.
3. Thinking vs. Feeling - measuring the tendency towards logic and reason or subjective emotional responses
4. Judgment vs. Perception - measuring the degree that an individual uses thinking and feeling or sensing and intuition.

Several studies on personality linked with social media have been conducted since much of our social interaction has migrated towards these public forums. This huge surge in data has allowed for acquiring personality and language data on individuals to become much more feasible and much less expensive. Related work in the field has been done primarily in relation to a user's Facebook and Twitter likes and social groups, however there are some studies focusing on use of language alone.

Anthony Ma and Gus Liu of Stanford University collected excerpts from author's works and their corresponding MBP Type and were able to achieve 28% accuracy in the baseline Bag of Words Feed-Forward Neural Network and a 37% accuracy with a Recurring Neural Network with Long Short Term Memory. The most related work belongs to Yilun Wang, also from the Department of Computer Science at Stanford, who used Bag of Words Techniques and created a model with the average AUC for 3 concatenated models across the 4 dimensions was 0.661. However, I believe implementing deep learning techniques with natural language processing along with systematic hyperparameter grid search techniques could improve on these results.

Problem Statement

The problem that I am setting out to solve is how to determine an individual's personality traits based on their use of language. If we can accurately predict one of the most fundamental aspects of a person's behavior and uniqueness based off the language they use, our ability to communicate information to that individual will be drastically improved. The internet is designed based on information that is already programmed into the web page itself; for example, administrators see a website much differently as a new customer or even a logged in user and are determined prior to visiting the webpage. This poses a difficult design problem for web designers to incorporate designs that maximize the profit or usefulness to their intended audience rather than communicating information or value to a person on an individual basis. If we can predict learning style or how an individual will react to their environment, then we can better customize the learning experience to their preference.

Quantifying personality has been done for us with the Myers Brigg Personality Index 4 letter code. These will be further broken down into their 4 features of a single letter with only 2 options, making it a binary choice and easy to encode the data. In addition, it will also allow us to train the weights of determining individual features of personality in a more focused way. NLP allows algorithms to extract meaning from text whether from word count, frequency, and even usage of words in a quantifiable and measurable way. Training a neural network on language use and their corresponding personality feature labels allows us to measure the AUC. Area under the ROC curve is used to ensure the proper binary classification when it comes to specificity and sensitivity. This will help better quantify individual differences in each of the 4 personality features.

Datasets and Inputs

The dataset will be taken from Kaggle and contains 8,600 users with 50 recent comments each on Kaggle and their corresponding personality type. This was user generated data from the Kaggle website and offers the most labeled personality data connected to their text data (comments) of what I could find online.

The dataset seems to be quite skewed according to other Kaggle user's visualizations, so several techniques will be used to counterbalance it. First of all, I will be using AUC as my evaluation metric and I may experiment with Kappa. I could also implement SMOTE and/or anomaly detection in order add synthetic data in personality types that are proportionately underrepresented (compared to the general population) in this dataset and/or remove outliers from it. I can experiment with different models, algorithms, and hyperparameters to help reduce the likelihood of my model learning to classify personalities that are overrepresented in the dataset. This will have to be done in each of the 4 personality dimensions as well as the overall personality prediction.

Another dataset I am considering is from openpsychometrics.org, This dataset of 1400 participants has data based on the Big 5 personality test, MBPI, and open ended sentences that needed to be completing. Having another training set to cross reference the model may help, but also may help in the training of a CNN where I can take the weights and use them on the Kaggle dataset. This may help remove some of the bias in the model that may result, since all participants in the Kaggle database have similar interests and occupations such that they participated in the Kaggle study in the first place.

Solution Statement

The solution will be reached by training a CNN to predict Myers Brigg's Personality Types and characteristics based on user's text. I will experiment between personality Type predictions (resulting in a prediction of 1 out of 16 potential personality types) and individual characteristics. There is some evidence to suggest that the personality types have overlay between characteristics (Citation). Perhaps more valuable than a potentially stereotyped personality archetype is to isolate individual characteristics, like extrovert vs introvert, in order to train the model.

Once the 4 models are completed, once for each characteristic in the MBPI, there may be higher accuracy in predicting the overall types by combining them than attempting to train a CNN to predict all 4 aspects at once. Even if there is not, it may help better understand the differences between many unique types introverts and extroverts; it is hard to believe that all people fall into 16 personality types that can predict behavior. In addition, separating the predictions can help find patterns between the characteristics. The goal is to have the highest average AUC for the overall model, with a potential consolation of predicting certain personality features over the benchmark model.

Benchmark Model

The benchmark model I will be comparing to is the one used by Simon Wang in his paper Understanding Personality Through Social Media. However, I will be using the Kaggle dataset, since I do not have access to the same data. Here, Wang uses NLP to create features like Parts of Speech, Weighted average word vectors, and bag of n-gram. He uses these features to predict individual personality features for each personality trait. He used over 90,000 twitter users' most recent 200 tweets for his dataset. With the best average AUC (after combining parts of speech, n-grams, and word vector, was 0.661. Wang used a Logistic Regression model with 10-fold cross-validation, although Random Forrest and SVM performed similarly well. Using CNN's especially RNN's should help understand the text and individual nuances of personality better than a n-gram and Bag of Words techniques. Therefore, I will use these simple models to determine a benchmark model on this particular dataset and hopefully achieve similar results. Then, I will use CNN's to compare the results.

There have been several other studies in recent years that show the potential of using MLP's, RNN, and CNN's. One, uses Neural Networks to predict Myers Brigg's Personality types of authors based on their use of language in books. However, this dataset is much different than the one I am planning to use and compares different NN architectures in order to compare their accuracy. Although I will not be using it as my benchmark, I will certainly use some learnings and suggestions for future improvement in this work. Finally, the most recent study done on this topic I can find compares social media text with personality type and features called Personality Prediction for Facebook Users. This uses 2 datasets of 250 and 150 facebook users respectively, much smaller sets than I would like to use, but achieves and accuracy of about 79.49% and up to 93.33% at determining extroversion on their manually collected set of 150 users, which is certainly something to aim for.

Evaluation Metrics

The benchmark and solution model will be evaluated primarily based on classification accuracy using AUC. Wang uses AUC in order to quantify and measure accuracy of a model. Since the distribution of personalities within the dataset is skewed in both our datasets, this will be a good evaluation metric to use. He also measured accuracy by comparing different models based on the features mentioned above based on AUC, not only breaking them down into dichotomous features (Sensing and Intuitive, Extrovert and Introvert), but also by focusing on features of the language. His best individual feature was average word vectors with an AUC of 0.651, which represents this model's ability to predict all 4 of the dichotomous personality traits.

Project Design

First, preprocessing of the data to remove bias from references to personality types, as well as other NLP techniques such as removing spaces, tagging parts of speech, and word vectorization will be done. Labeling of personality types will also have to be preprocessed since I want to break apart the personality features in order to fine tune the loss function better than training for 16 features. This will also help understand the predictive abilities of models on certain personality features rather than trying to generalize them all at once. Next, I will run CNN's and hyperparameter cross validation in order to systematically tune my model to get the best results. Finally, the results can be combined into another NN to set weights on each personality feature and hopefully improve overall accuracy.

In order to compare overall accuracy, I will train a CNN to predict the full personality type with all 4 features and compare those results to the individually based predictions, and the CNN of the 4 predictions to produce the entire personality prediction. With these comparisons, I will be able to visualize the predictive power of several models and individual aspects of personality based on use of language.