Machine Learning Proposal

Jean-Michel Mailloux-Huberdeau
j_m789@txstate.edu

Gridley, Garrett
gbg15@txstate.edu

February 2018

1 INTRODUCTION

The Meyers-Briggs Type Indicator (MBTI for short) is a self-reported test that indicates psychological profiles about decision making and perception of the world. The Meyers-Briggs test is known to have inaccuracies, due to the question being written to fit the different categories they developed and attributed to people as opposed to the other way around. Our goal is to evaluate the validity of the test by attempting to predict an individuals MBTI based on their online language usage and style. That is, predict their MBTI purely off their online posts. The data set we are pulling from contains 50 posts each from 8600 peoples with known MBTI.

2 METHODOLOGY

A similar analysis has been done by Stanford using neural networks in addition to unsupervised learning to predict personality types by reading an individual's writing. We plan on using supervised learning in addition to some of the methods implemented by Stanford. We believe that the addition of our data sets known MBTI will lend towards us having a more accurate prediction rate than their 37%.

3 MILESTONES

1. **Organize And Study Dataset:**
   Previous studies of this type found that datasets with smaller data points were not learning properly and were instead simply predicting common MBTI types, and that pruning data points in order to balance the different personality type groups was needed. We will also have to work on a structure of the data as it is not in an easily manipulable format.

2. **Run Preliminary Training:**
   We will utilize supervised learning based on the given MBTI to begin training a neural net that can identify MBTI based on an individuals posts. It might be useful to use both supervised and unsupervised learning to visualize the difference in prediction accuracy.

3. **Attempt To Increase Accuracy of Model**
   Hopefully by this point we will have good preliminary results such that we can make efforts to increase the models overall accuracy. Again, our goal is to achieve an accuracy higher than that of Stanford's 37%.

4. **Apply Model To New Datasets**
   Assuming all milestones are met, it may

be wise to seek out other data sets that we can apply our model to in order to test further test the prediction model. This is a hopeful milestone, as opposed to something that we consider to be critical for the completion of the project.

4 REFERENCES

References to be properly formatted at a later date, currently in the order found.

1 https://web.stanford.edu/class/cs224n/reports/2736946.pdf

2 https://nlp.stanford.edu/courses/cs224n/2015/reports/6.pdf

3 https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS12/paper/viewFile/4417/4799