Neural Networks in Predicting Myers Brigg Personality Type From Writing Style

Anthony Ma
Department of Computer Science
Stanford University
Palo Alto, 94305
akma327@stanford.edu

Gus Liu
Department of Computer Science
Stanford University
Palo Alto, 94305
gusliu@stanford.edu

Abstract

Personality is the defining essence of an individual as it guides the way we think, act, and interpret external stimuli. Over the past century, aspects of personality have been studied from many angles whether through analyzing interpersonal relationships, team dynamics, and social networks or through works in neuroscience that reveal the biological underpinnings of personality traits. While many components of our personality remain consistent with time, behaviors are not as stable given they adapt to environmental situations and integrate habits that one accumulates throughout their life. Understanding the underlying essence of a person amidst the noise of behavior is a very highly sought out problem. Many studies have aimed to predict personality by analyzing patterns in ones behavior, pictures, and even handwriting. The brain regions that encode for various personality traits are often coupled with regions responsible with verbal and written communication. Furthermore, the advent of social media and an increasingly connected online community makes personalized textual data increasingly available. In this study, we hypothesize that an individuals writing style is largely coupled with their personality traits and present a deep learning model to predict Myers Briggs Personality Type through textual data from books. Developing an accurate model and opening this question of research would have significant implications in the business intelligence, relationship compatability analysis, and other fields in sociology.

1 Introduction

Personality is regarded as one of the most influential research topics in psychology because it is predictive of many consequential outcomes such as mental and physical health, quality of interpersonal relationships, career fit and satisfaction, workplace performance, and overall wellbeing (Li et al. 2014). It is widely known that personality traits such as extraversion, conscientiousness, and neuroticism are relatively consistent throughout ones life. However, the ways in which our behaviors are expressed via words and action are not always determined by underlying personality traits and impulses alone; people effectively learn to modulate their behaviors to align with habits and external circumstances (Read et al. 2010). Many important decisions social dynamics and political decisions are also based on judging the personality of an individual whom one has not interacted much with personally. Overall, personality traits are highly influential in affecting our behavior, but reading another persons behaviors alone is not sufficient in making accurate prediction of their personality. The task becomes even more challenging when trying to make judgements based on written communication alone. Because, the world is relying much more heavily on text-based communication than face-to-face interactions, it is becoming increasingly important to develop models that can automatically and accurately read the essence of other individuals based on

writing alone. Fortunately, studies in neuroscience have revealed close mappings of brain regions responsible for personality traits such as extraversion and neuroticism as well as those that are linked to written communication (Adelstein et al. 2011). Given the highly interconnected nature of neurons, we have reason to believe that underlying patterns of personality can be extracted from written text.

Previous personality prediction models have focused on applying general machine learning techniques and neural networks to predicting the Big-Five personality traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism from social media posts (Schwartz et al. 2013). Other studies have incorporated computer vision techniques to predict personality traits from profile pictures, handwriting, and other image based data (Liu et al. 2016). While, utilizing textual and image based data from social media websites would be useful in predicting friendship dynamics in large social networks, they dont reveal the full spectrum of ones behavior. Writing from posts and tweets tend to be in the style of prose and has great variability which allows for better differentiation between say introverts and extraverts. Often times, the more challenging task is to pick up nuanced differences in individuality and personality traits from more formal and standardized writing styles such as essays, emails, or job applications. Furthermore, the studies that focus on the Big-Five traits tend to give a trait by trait picture of an individual, whereas Myers Briggs Personality Type (MBTI) tend to be associated to archetypes that are more easily comparable and have functional applications in predicting compatibility in vocation and relationships as well as behavioral tendencies.

In this work, we explored a variety of methods to address the personality prediction problem. We started by manually building a large corpus mapping excerpts from famous novels with MBTI types of authors. To gauge the difficulty of identifying MBTI from text, we clustered text segments based on word embedding similarities to determine whether there existed any non-uniform distribution of personality types. This provides a good starting frame of reference to understand how subtle personality traits are when hidden in written data. We then implemented a bag of words feed-forward neural network as a baseline to understand how simple models in deep-learning can provide insight of hidden personality features. Finally, we delve into a more complex long-short term memory based recurrent neural network and aim to build a more generalizable system that can incorporate meaning of writing to determine overall personality types.

2 Background

2.1 Personality and Myers-Briggs

The Myers-Briggs Type Indicator (MBTI) is based on Carl Jungs theory of psychological types which states that random variation in behavior is accounted for by the way people use judgement and perception. There are 16 personality types across four dimensions. Extraversion (E) vs Introversion (I) is a measure of how much an individual prefers their outer or inner world. Sensing (S) vs Intuition (N) differentiates those that process information through the five senses versus impressions through patterns. Thinking (T) vs Feeling (F) is a measure of preference for objective principles and facts versus weighing the points of view of others. Finally, Judging (J) vs Perceiving (P) differentiates those that prefer planned and ordered life versus flexible and spontaneous. Note that these measures are not binary but rather on a continuum.

MBTI	Archetype	Traits	% of population
INTJ	Architect	Imaginative and strategic thinkers, with a plan for everything.	2.1
INTP	Logician	Innovative inventors with an unquenchable thirst for knowledge.	3.3
ENTJ	Commander	Bold, imaginative and strong-willed leaders, always finding a way – or making one.	1.8
ENTP	Debater	Smart and curious thinkers who cannot resist an intellectual challenge.	3.2
INFJ	Advocate	Quiet and mystical, yet very inspiring and tireless idealists.	1.5
INFP	Mediator	Poetic, kind and altruistic people, always eager to help a good cause.	4.4
ENFJ	Protagonist	Charismatic and inspiring leaders, able to mesmerize their listeners.	2.5
ENFP	Campaigner	Enthusiastic, creative and sociable free spirits, who can always find a reason to smile.	8.1
ISTJ	Logistician	Practical and fact-minded individuals, whose reliability cannot be doubted.	11.6
ISFJ	Defender	Very dedicated and warm protectors, always ready to defend their loved ones.	13.8
ESTJ	Executive	Excellent administrators, unsurpassed at managing things – or people.	8.7
ESFJ	Consul	Extraordinarily caring, social and popular people, always eager to help.	12.3
ISTP	Virtuoso	Bold and practical experimenters, masters of all kinds of tools.	5.4
ISFP	Adventurer	Flexible and charming artists, always ready to explore and experience something new.	8.8
ESTP	Entrepreneur	Smart, energetic and very perceptive people, who truly enjoy living on the edge.	4.3
ESFP	Entertainer	Spontaneous, energetic and enthusiastic people – life is never boring around them.	8.5

2.2 Neural Network and Supervised/Unsupervised Machine Learning Background

In this section, we justify our neural network-based approach and compare it with popular machine learning methods. We first establish that there is currently scarce existing research that has found the most powerful features for predicting personality from text. From this lack of feature engineering, it is already difficult to use standard machine learning methods to perform this difficult task. Next, it is certainly conceivable to construct a massive dataset for this task with the appropriate amount of effort and time spent on data collection, allowing neural networks the bandwidth to learn the important and relevant features while avoiding overfitting. Now, we consider the models themselves. A recurrent neural network takes in the sequential nature of the data with the context from multiple previous time steps fed to the next timestep. We hypothesize that sentence progression, structure, and flow is as important as content when determining personality type. For example, an introvert may write the same thing in terms of word choice as an extravert but with very different tone. A concrete example is that an introvert may write, It is possible, I think. On the other hand, an extravert may write I think it it is possible. The content of each sentence is the same, something that even our baseline bag of words model treats as the same, whereas an LSTM has the capability to deduce that the first is more representative of an introvert and the second an extravert. Not only can an LSTM learn what long term dependencies to incorporate, it can conversely decide what information to forget and ignore at each time step. Given our dataset of books written by celebrated authors, it is clear that our authors chose each sentence and paragraph structure with great care. We investigate whether those choices are reflective of personalities.

3 Data

3.1 Overview

To train our model, we needed to obtain a data set of text segments associated with the Myers-Briggs personality type of the author. To manually build this dataset, we first generated a mapping between famous authors and MBTI by using Google API, MBTIDatabase, and BookRiot. After identifying the writers, we searched for a book from ten authors for each personality type. Doing so would help us balance our data such that there is enough variability within a personality type. The books were obtained from free book repositories such as The Gutenberg Project, Stanford Online Library, and EBookCentral. We then converted these books from PDF format to .txt files using online conversion engines. In total, we were able to manually curate over 750,000 labeled sentences.

3.2 Preprocessing

The .txt files containing the raw book contents were parsed and formatted to remove sentences written in the table of contents, index, publishing detail, and any information that was not written by the author. The entire text was then split upon punctuations to generate sentences. We also removed numbers, non-English words, urls, and extraneous information contained in PDF versions. Next, we generated data points that were chunks of five sentences - a length that is typical of a short paragraph that contains appropriate contextual information. Chunk sizes that are too large would be less feasible for personality prediction due to vanishing gradient problems. Excessively small chunks would omit important context that is important in capturing features of ones personality. The chunk size remains as a hyperparameter that we would hope to optimize carefully. We represented the sentences as one-hot vectors corresponding to the words, from which we obtained their GloVe vectors using pre-trained embeddings. Similarly, we represented our output vectors as one-hot vectors of length 16 with a value of 1 for the class and 0s elsewhere.

3.3 Dataset Issues

Initially we ran our deep-learning model on the full dataset, but found that deep-learning model was fitting to the distribution of personality types in our training data rather than learning. In other words it always predicted the most common MBTI types while have near zero predictions in the less common ones. It turns out that each book varied greatly in the number of post-processed sentences, and therefore not every personality type had the same number of effective data points. We decided to balance our data set by determining the minimum number of datapoints across any category and then randomly picking this number of points from all 16 MBTI types to build a balanced and shuffled data set. After pruning, our data set had about 50,000 chunks and a total of approximately 250,000 sentences.

4 Models

4.1 Unsupervised Clustering with SVD

To gain a sense of how subtle personality type is encoded into writing, we perform unsupervised machine learning to cluster sentences and see whether certain clusters have non-uniform distribution of certain Myers Briggs classification. Every word of a five-sentence chunk was converted to a 50-dimensional glove vector. These vectors were averaged to generate a single 1x50 vector that represents the content of the sentence. SVD was performed to extract the top two singular vectors in which we were then able to plot the averaged glove vectors across a 2-dimensional map. The distribution of personality types serve as a null model reference to gain a sense of how subtle personality traits are embedded in writing.

4.2 Bag of words Feed-Forward Neural Network

We implemented a bag of words feed-forward neural network as a baseline to the MBTI prediction problem. Effectively, each word of a five-sentence chunk was converted to a 50-dimensional glove word embedding, averaged together and fed into a standard feed-forward model. The weights and biases were defined by a 50 x 16 and 1 x 16 matrix respectively given there were 16 possible personality classifications. Weights were initialized with a Xavier initializer, while the biases were initialized with zeros. This model was run over 100 epochs. We used a gradient descent optimizer to optimize the loss with a learning rate of 0.001.

4.3 RNN with LSTM

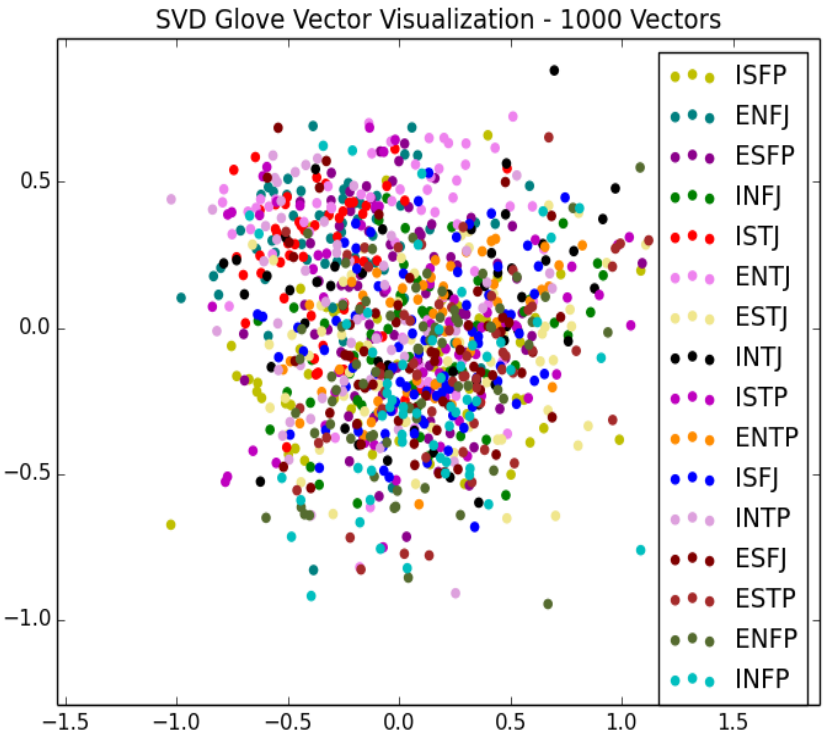
The main model presented in this paper is a RNN with LSTM. The five sentence chunks were processed into array of words. This array was either truncated or padded with empty strings to be at predefined max sequence length = 120, and fed into a recurrent neural network with weights and biases defined by a 50 x 16 and 1 x 16 matrix respectively. Batch size was set to be 64, and the model was run over 1700 epochs. Our input was of shape 32 x 120 x 50 and our output was of shape 32 x 16. We extracted the last output by timestep and predicted using that tensor.

4.4 Model Tuning

Initially, we implemented our model with a max sequence length of 240 with batch sizes of 32. After training the model with these parameters, we found that performance was poor and training was very slow. We hypothesized that our model was suffering from a vanishing gradient and that our batch sizes were too small, and our VM had sufficient memory to handle a larger batch. Thus, we tuned our parameters by decreasing the maximum sequence length and increasing the batch size, which improved both performance and training speed. Another parameter we heavily tuned was the number of epochs. Initially, we started with only 100 epochs, but we realized that our error was not quite converging yet, and so we greatly increased the number of epochs to allow the model more iterations to learn parameters from the data. This was different than in the baseline, where our error rate leveled off after around 60 or so epochs, and so only 100 epochs were necessary in that case.

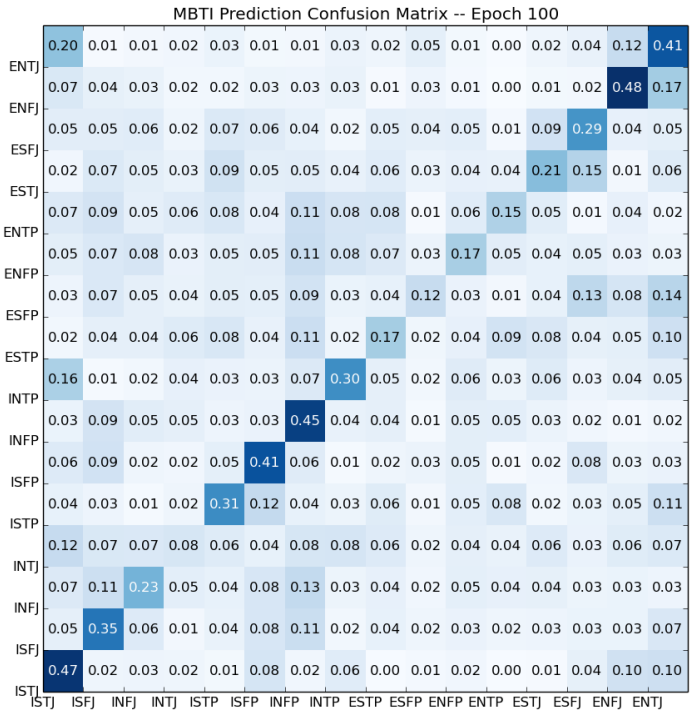
5 Results

5.1 SVD Visualization



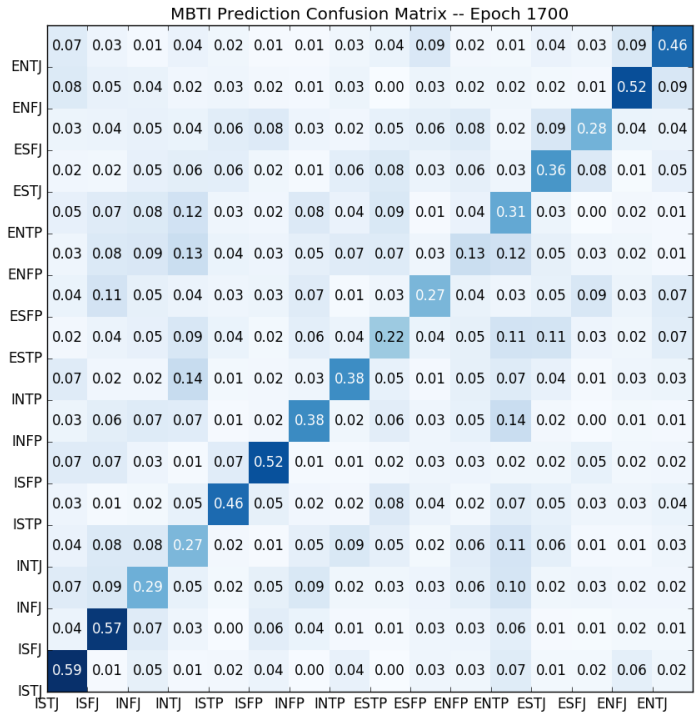
From our SVD analysis, it seems that there are no cluster formations or identifiable patterns in the average GloVe vectors for each data point. This indicates to us that our task is nontrivial, as a low-dimensional representation of the data yields no meaningful interpretations. This further supports the need to use a neural network to capture the interactions between the vectors at their high dimensions.

5.2 Baseline

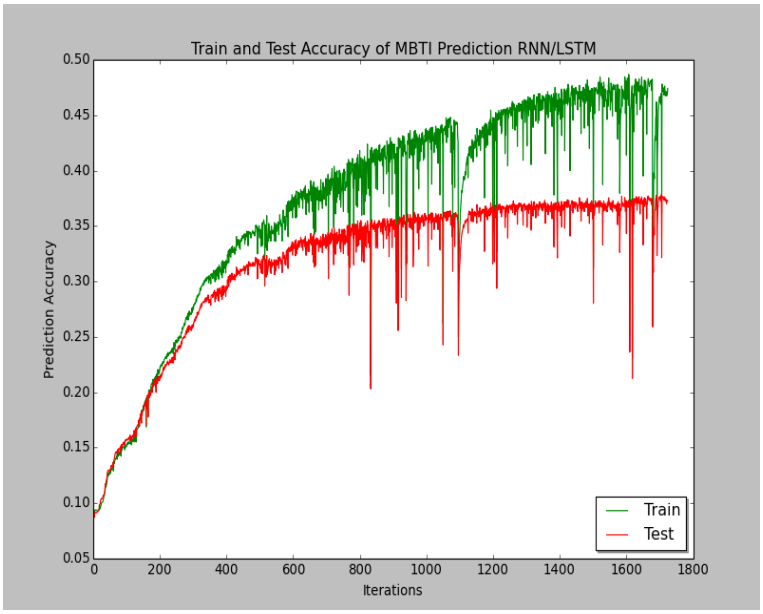


Our baseline BOW feed forward neural network performed quite well with an accuracy of around 28%. As we can see from the confusion matrix where the y-axis is the true labels and the x-axis is the predicted classes, the diagonal of true positives is heavily weighted in comparison to every other value. We scaled each row by the sum of the row, therefore calculating the proportion of each class we predicted for each true label. We can see that the model has difficulty distinguishing the INTJ and ESFP personality types with a lot of false negatives and false negatives. Moreover, there is a good amount of noise in the non-diagonal entries. With such a simple model, we can see that we have already captured a significant portion of the distinguishing features between personality types in writing. The advantages of this sort of approach is that the model is simple and easy to implement, as well as relatively quick to train and tune. However, the disadvantages of such a simple model are clearly that it has limited capabilities, and it does not take into account the sequential nature of sentences.

5.3 RNN with LSTM



(a) Confusion matrix at epoch 1700



(b) Training and test accuracy vs epochs

Our best RNN model with LSTM significantly outperformed our feed forward baseline network, with an accuracy of 37%. We can see in the confusion matrix that the noise around the diagonal is significantly reduced. Interestingly, we found that for the first few hundred epochs, a diagonal does not appear to materialize in the confusion matrix. In fact, the model seemed to only predict a few classes and ignore the rest, resulting in a very low accuracy. After several hundred epochs, we begin to see a heavily weighted diagonal appear that is definitively better than the baseline. We conclude that the LSTM learns at a much slower rate than our baseline, needing more passes over the data to effectively learn distinguishing features. However, this allows for the potential of more information to be encoded in the parameters. We also notice that our model begins to overfit the training data, calling for the need for future regularization and dropout implementation.

6 Discussion

We trained a recurrent neural network with LSTM units for predicting MBTI personality type using excerpts from novels. Using only five sentences of text, this model is able to predict the most likely personality type of a writer. Our model does well at predicting overall personality trends but can use more work in developing more sophisticated neural networks, well rounded evaluation metrics, and expansive data sets. This work adds upon existing deep-learning models that predict other features of personality as well as the work that has been done through supervised machine learning approach. It is possible, that the integration of methods would lead to the most accurate system in personality prediction which could lead to widespread applications in social network theory and psychology.

7 Conclusion

Our experiments show that neural networks achieve a significant level of effectiveness predicting personality from written text. The best model of a dynamic RNN with LSTM predicted with an accuracy up to 37%. We achieved these results after some careful tuning of the maximum sequence length to address the vanishing gradient problem. We postulate that we have built a solid framework and exploration into this task that can be built upon and extended with several ideas that we will discuss in the next section.

8 Future Steps

While our models demonstrate that neural networks can predict MBTI from written text effectively, it has yet to be seen if more complex models can achieve better results on the same task. For example, introducing bidirectionality to our RNNs can incorporate future information along with past information for any given token, which could improve our results even further. We also hypothesize that a more complex loss function could yield better training. Intuitively, it does not make sense to treat all classifications as disjoint and penalize the same way. In particular, a misclassification of INFJ as ISFJ should not be penalized as much as a misclassification of INFJ as ESTP. Thus, we propose exploration of a more suitable weighted loss function by personality dimension for our task, one that possibly computes cross-entropy by letter and sums up the individual losses. This would allow our model to learn more nuanced differences between personality categories and perform more finely grained steps. A more comprehensive hyperparameter grid search could be useful as well, given the appropriate time and resources to do so. The hyperparameters we think could still be optimized are chunk length and hidden size. In addition, we could implement regularization and dropout rate to avoid overfitting our dataset. Finally, our dataset could be expanded to include many more sentences for each personality type. With a large enough dataset, it is possible to segment the data by genre, time period, or topic so that we could reduce the number of variability that causes noise in the data. By doing so, we can ensure that our model is learning personality with high probability as opposed to other factors.

9 References

[1] Adelstein, Jonathan S. et al. Personality Is Reflected in the Brains Intrinsic Functional Architecture. Ed. Mitchell Valdes-Sosa. PLoS ONE 6.11 (2011): e27633. PMC. Web. 19 Mar. 2017.

[2] Champa, H. N., and Dr. K R Anandakumar. "Artificial Neural Network for Human Behavior Prediction through Handwriting Analysis." International Journal of Computer Applications 2.2 (2010): 36-41. Web.

[3] Kalghatgi MP, Ramannavar M, Sidnal N (2015) A Neural Network Approach to Personality Prediction based on the Big-Five Model. International Journal of Innovative Research in Advanced Engineering (IJIRAE)

[4] Li L, Li A, Hao B, Guan Z, Zhu T (2014) Predicting Active Users' Personality Based on Micro-Blogging Behaviors. PLoS ONE 9(1): e84997. doi:10.1371/journal.pone.0084997

[5] Liu L, Pietro D, Samani Z, Moghaddam M, Ungar L. Analyzing Personality through Social Media Profile Picture Choice. AAAI Digital Library (2016) Web.

[6] Read, Stephen J., Brian M. Monroe, Aaron L. Brownstein, Yu Yang, Gurveen Chopra, and Lynn C. Miller. "A Neural Network Model of the Structure and Dynamics of Human Personality." Psychological Review 117.1 (2010): 61-92. Web.

[7] Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. (2013) Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE 8(9): e73791. doi:10.1371/journal.pone.0073791

[8] Scott, Kate. "The Myers-Briggs Types of 101 Famous Authors." BOOK RIOT. N.p., 11 Nov. 2015. Web. 19 Mar. 2017.

[9] "The MBTI Database — Personality Profiling." The MBTI Database — Personality Profiling. N.p., n.d. Web. 19 Mar. 2017.