Predicting Personality From Writing Style

Damayanti Sengupta **Sayak Ghosh** **Sagar Gurtu**
Department of Computer Science
Stony Brook University, New York
damayanti.sengupta@stonybrook.edu
sayak.ghosh@stonybrook.edu
sagar.gurtu@stonybrook.edu

Abstract

This work falls under the domain of psycholinguistics and analyzes the impact of personality on writing style. We have attempted to examine the relationship between language semantics and traits of an individual and predict personality temperaments using features like subject matter, vocabulary usage, and sentence construction. We have used a mapping between literary texts by published authors and their Myers-Briggs Type Indicator to learn neural network models that can classify personality from textual input.

1 Introduction

1.1 Objective

The focus of this project is the link between facets of personality (i.e. cognitive processes underlying someone's behavior, ideals, attitude and worldview), and the expressions they use to communicate with the external world. Research suggests that language is inextricably tied to the mental processes of a human being, and this project aims to investigate whether the claim holds true for patterns found in a person's writing style. Personality categorization and detection has found use in many domains like recommendation engines, compatibility matching platforms, mental health diagnosis, forensic and fraud analysis, human resources management, etc. Since personality is an attribute of the human mind, it is very hard to measure and understand. Our objective is to classify the personality type of an author under Myers Briggs Type Indicator using their published work as dataset. The input of the model will be text, which has been obtained from books, plays, poetry, journals, etc. and the output will be one of the 16 personality classes of MBTI.

1.2 Related Work

A majority of existing approaches for the problem of computational personality modelling deal with a different but comparably popular metric known as the Big Five Model. The Big Five Model quantifies personality traits in five broad categories: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These five descriptors encapsulate a range of words used to categorize human behavior, and capture how susceptible to change, participation in certain activities/hobbies, and reasonable they are. Also, a majority of the research as of late has been focused on fetching the textual data for this modelling from social media, as then the model can be directly (or with minor changes) utilized in applications that target users of the same forums from which the data has been collected. Some of the earlier work include using convolutional neural networks to derive personality traits based on Big Five by Majumder et al. (2017). Plank and Hovy (2015) used a logistic regression classifier to detect personality dimensions on Twitter. Another approach by Liu et al. (2017) uses feed-forward neural network called 'Character to Word to Sentence for Personality Trait (C2W2S4PT)' on twitter dataset. Apart from textual analysis, others have used features like profile pictures (Liu et al., 2016) and handwriting analysis (Champa and AnandaKumar, 2010) for predicting personality traits.

1.3 Issues and Proposed Solutions

One major issue in dealing with personality prediction seems to be limited availability of reliable dataset with authentic trait annotations which is hard to obtain. Almost all the previous works derive data from social media posts like tweets which may not be fully reflective of a persons psyche and

might even project a person's alter ego to fit a certain public image since users tend to flaunt and imitate herd instinct. Further, they collected tweets from users who have self-reported their personality types, thus suffering from selection bias and possibility of being dishonest or inaccurate. Moreover, they may provide a broad perspective of a certain personality trait but might not capture subtle differences and overlapping of multiple character attributes. Additional challenges like presence of emoticons, slang terms etc. requires more effort in the cleaning phase. Therefore, in our work, we have used text from published sources only which will cover a range of different genres.

In our opinion, the Big Five personality model does not offer as many trait dichotomies as the MBTI Indicator does. So, for our model, we have used MBTI personality types since these are already evaluated and available for all the famous personalities we have included in our dataset (MBTI Database), and we can avoid the issue of inaccurate self-reported traits. In these papers, we've also noted that certain personality types were harder to detect due to rarity and relatively less annotations for these types. Therefore, we used a uniform distribution of the 16 MBTI classes in our dataset.

Another drawback present in previous approaches was the need for crafting of domain-specific features which required extensive knowledge of psychology and stylometry. We removed the need for manually designing features by using low-dimensional dense vectors such as word embeddings (either word2Vec or GloVe). Further, a more complex network which models attention over chunks and focuses on relevant contexts while predicting specific personality traits could remove noise.

1.4 Model

We have used a logistic regression classifier and a feed forward neural network as the baseline models. For our dataset, we collated a list of authors and mapped them to their MBTI from the MBTI Database and Google. We then gathered their respective published articles, books, etc. which were freely available on the Internet and chunked the text to specified sizes which we annotated with the author's MBTI. We evaluated these models by computing the accuracy of correct MBTI predictions on the test set.

1.5 Outcomes

The main outcomes of this project are:

1. We implemented a neural model as part of an end-to-end system that outputs a MBTI classification for an author based on a text written by them given as input.

2. We extended the neural model by extending it in several different ways. We implemented a single-layer LSTM architecture, two different bi-directional RNN models (LSTM and GRU), and a bi-directional LSTM word-to-sentence model.

3. Our evaluation shows that our machine learning baseline (four logistic classifiers with word embedding as features) performs much better than majority class classifier (i.e. assuming a uniform distribution and randomly assigning a class from that distribution). The feed-forward neural network with one-layer improves upon that performance accuracy, with two-layers increasing that improvement. Finally, we test several different recurrent neural network architectures that do not converge in the training time we have used to compare all the methods (500 training epochs), but which achieves comparable accuracy, failing to outdo the simpler and faster feed-forward model.

4. Our observation was that the LSTM-based models were slow to learn, but also encoded the sequential nature of the word embeddings that we used as input, constructing a comprehensive representation over many time steps. In our opinion, the LSTM-based models could not outperform the simple feed forward neural model due to the fact that our chunks were long sequences (with a maximum length of 120 words), and this could lead to forgetting past information down the line for the longer chunks. Hence, we need to add an attention layer that can focus on different segments of the chunk when deciding between the 16 classes. In this way, the model can learn which personality attributes can be inferred from which parts of the chunk.

2 Task

The task definition is: given a textual input, output the most probable MBTI class. One of the

key challenges that need to be addressed is how to capture the nuances of the writing style in the text. That is, how linguistic features like sentence formation, lexicons as well as influence of subjectivity and native culture on the written text can be learnt in order to predict an individual's personality. Some of the state-of-the-art models use approaches like LR classifier (Plank and Hovy, 2015), SVM and multi layer perceptron (Gjurkovic and Snajder, 2018) and Bi-RNN C2W2S4PT (Liu et al., 2017)

2.1 Baseline Model

We have implemented a two-layer feed-forward neural network, which takes the word embeddings (averaged over the sequence of chunks) as an input, and returns a prediction of the personality class as an output.

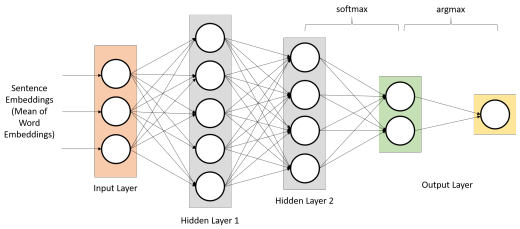

Figure 1: Feed-Forward Neural Network

2.2 Issues

Our baseline model is already performing at quite a sophisticated level compared to simpler approaches involving machine learning which perform classification tasks (decision trees, SVMs, boosting methods, etc. using features). The main issues with the approach are as follows:

- The feed-forward network does not capture any non-linearity that may exist in the relationships between the implicit features encoded in the text

- The network uses a sentence representation that is made by averaging the word embeddings together. While this is an adequate representation, not all words are relevant for the task, and it aggregates information where we may benefit from the atomic representations of the words fed as a sequence to the network.

3 Our Approach

3.1 Chunks Represent Stylistic Patterns Over Document

The key idea encapsulated in our dataset processing and transformation to the input fed into our model is the principle that the stylistic patterns present in a document can be well-represented by a randomly selected chunk with a length above a certain threshold. In our project, we have defined a chunk as a collection of five consecutive sentences. We believe that five sentences are enough to capture the flow, the linguistic quirks, the usage of words and the vocabulary that can be generalized over the entire text for that particular author.

3.2 Sentence Representations from Word Representations

Our neural models are trained on the word embeddings of the constituent words in the chunk. The idea is that sentences are constitutional representations of the word sequences. We believe that the subtle differences in writing style can be expressed within those sentence representations.

3.3 Personality Detection from Style Inference based on Sentence

Finally, our sentence representation and the final personality class can be modeled using complex non-linear relationships that can be arrived at by using neural models to learn them.

3.4 Implementation Details

For our final model we have used a Recurrent Neural Network with LSTM. We were motivated to use recurrent neural network to derive sentence embeddings as they capture the sequence of words and as a result , stylistic and syntactic features embedded in a certain ordering of words. But the problem is that Recurrent neural network by itself does not capture long term dependencies. This problem is amplified while using longer sentences as input. As our input is a chunk containing 5 sentences its inherently bound to be long. For this reason we used LSTM cells in conjunction with RNN to mitigate the issue.

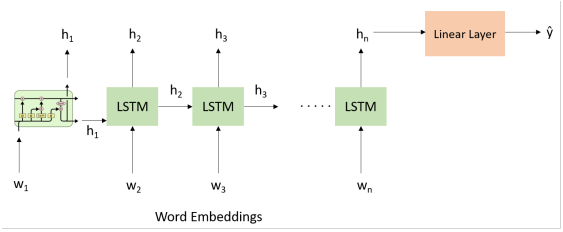

Figure 2: LSTM Model

The model is defined as below:

- Input - Word embeddings w_1,w_2,...w_n.
  n is the maximum length of the sequence - 120
- If the length of the sequence is less than 120, it is padded with white spaces in order to maintain a fixed length for the sequence. If it's more than 120 then the sentence is trimmed to the maximum sequence length.
- The final output i.e. h_n is the final sentence embedding which is passed through one hidden layer to derive the final output vector $\hat{y}$ after passing it through a softmax layer.
- The argmax of the softmax layer is taken as the output class.The trained model had an accuracy **30%**
  Hyper parameters:
  - LSTM cell size: 50
  - Learning rate: 0.05
  - Output layer hidden size: 50
  - Word Embedding size: 50

4 Evaluation

The purpose of our evaluation is to measure how well our model performed for the given task and the correctness of our model. The questions here are: what is the baseline accuracy, how well does the model beat the baseline and the state-of-the-art models and what accuracy did we achieve with our model.

4.1 Dataset Details

We considered 10 authors for each personality type (16 classes), that amounts to 160 authors in total. For each of these authors, we looked up their publications and processed the text from them in order to create training instances. We used 47,025 chunks of text with uniformly distributed MBTI annotations with each chunk consisting of 5 sentences.

4.2 Evaluation Measures

We segmented the dataset into training, dev and testing sets and used the testing sets to validate our model and derive our accuracy scores.

4.3 Baselines

We used the following models as baselines:

- **Logistic Regression Model**
  Logistic Regression is a machine learning model for classifying data points across regression lines. The regression model is a linear function of indicator variables, with the dependant variable being a binary variable.

  For our purpose we modelled our baseline as 4 separate binary classifiers which classifies the following respectively
  - Introversion/Extraversion
  - Intuition/Sensing
  - Feeling/Thinking
  - Perception/Judgement

  We derive the final MBTI personality type from the combination of the 4 outputs.
  The features used are simply the sentence embedding of the training instances. Sentence embeddings are a mean of the word embeddings of the words contained in the chunk. The trained model had an accuracy **19.75%**
  Hyper parameters:
  - C (Inverse of regularization strength): 10^5
  - Word Embedding size: 100
- **Feed Forward Neural Network**
  Feed forward neural network also known as MLP(Multi Layered Perceptron) is the simplest form of artificial neural network where the information flow through the hidden layers to the output is only in one direction ie. forward. The network defines a mapping $y = f(x|\theta)$ and learns the value of θ that results in the best approximation of the function.
  We used this model as our baseline as we thought that it would better capture the non linearity of the features, inherent in our problem in comparison to a linear function as in logistic regression.
  The model is illustrated in Figure 1.
  Model Parameters:

- Input Layer: Sentence Embeddings derived as a mean of the word embedding of the words contained in the sentence.
- Hidden Layer 1: A hidden layer with 100 nodes
- Hidden Layer 2: A hidden layer with 50 nodes
- Softmax Layer: A vector of size 16 i.e the number of MBTI types.
- Output class: Argmax of the softmax layer
- Loss Function: Cross entropy

Hyper parameters:

- Hidden Layer 1 size: 100 nodes
- Hidden Layer 2 size: 50 nodes
- Learning rate: 0.005
- Number of epochs: 300

The model had an accuracy of **32.5%** with the above hyper parameters.

4.4 Results

Following are the results of our models:

Model	Accuracy	No. of Epochs
Logistic Regression	19.50%	N.A.
Feed-forward (1 layer)	26%	200
Feed-forward (2 layers)	32.50%	300
RNN (with LSTM)	30%	200
bi-RNN (with LSTM)	21%	150
bi-RNN (with LSTM) + Feed-Forward (2	17%	100
RNN (with GRU)	22%	150

Table 1: Accuracy of different models

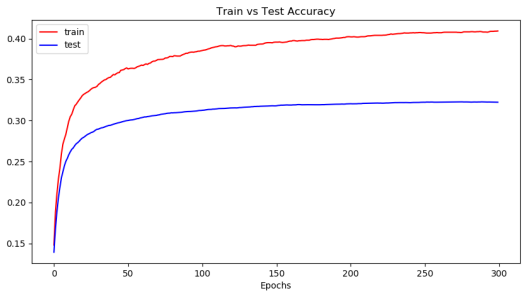

Figure 3: Train-Test Accuracy for Feed-Forward Baseline

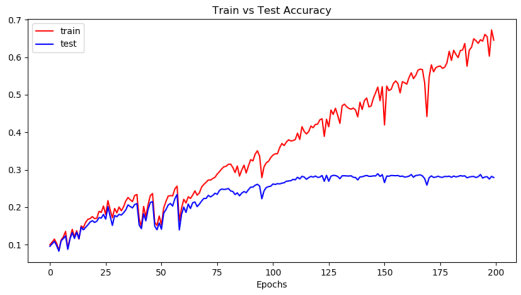

Figure 4: Train-Test Accuracy for LSTM

4.5 Analysis

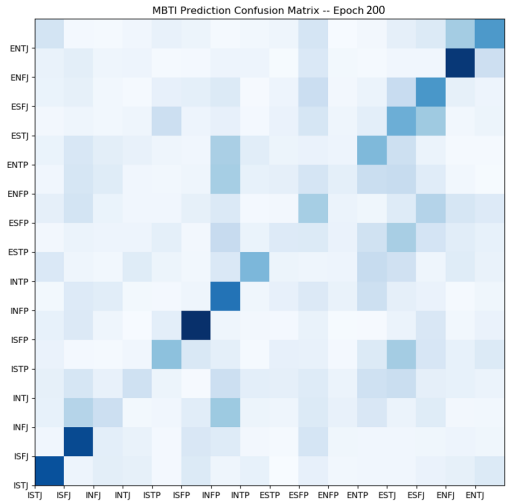

Figure 5: LSTM Model Confusion Matrix for MBTI classes

1. For a short number of epochs, the two-layer feed forward neural network is the clear winner as it is fast to train and flexible enough to capture non-linearity. The feed-forward model plateaus after 100 epochs, while the loss does not converge for the LSTM-based models.

2. The confusion matrix shows a strong diagonal, but we can see that our model is only confident about certain classes. This could imply that it is learning certain stylistic patterns better than others. These are ISTJ, ISFJ, INFP, INTP, ENFJ. This implies that the model is learning introverted personality features faster.

3. Adding attention, dropout layers, and regularization for our parameters can combat the problem of forgetting past information over

the long sequence, as well as overfitting due to the small amount of training data available.

4.6 Code

Our code can be viewed here at Google Drive .

5 Conclusions

The first assessment we can derive from our efforts is the strong link between personality traits and writing patterns. These patterns involve not just the semantic and syntactic features of the text, but also the topics they address, the frequency of specific words, and other such lexical factors. We obviated the need for extensive domain knowledge, by not considering manually curated features, which also suffer from lack of full coverage of the complex, non-linear relationship between the text and the personality facets. We also implemented and compared several layered neural architectures, so we could encode stylistic features in different ways from our input of word embeddings.

Our baseline considered sentences to be an averaged representations of their granular embeddings, while our more sophisticated and complex architectures learnt this representation. Our basic RNN model took the words as a sequence of time-steps, and constructed an encoded representation of the chunk in that manner. Further, we implemented a compositional model that learnt a lot of implicit information due to the bi-directional RNN layers, and also built a feed-forward layer on top of that which took the single sentence representation and learnt how to classify that into the MBTI classes.

Based on the evaluation of the results of these experiments, we can conclude that even simple neural models are powerful enough to capture the complex relationship between personality and language.

Future work can also include efforts to extend the dataset so we can account for the noise generated due to genre-specific keywords and remove the influence of that from our models. Additionally, a more complex loss function that penalizes misclassifications differently (e.g. INTJ is close to INTP, vs INTJ is very far from ESFP) can help improve the models we have experimented with in this project.

References

HN Champa and Dr. KR AnandaKumar. 2010. Artificial neural network for human behavior prediction through handwriting analysis. *International Journal of Computer Applications*.

Matej Gjurkovic and Jan Snajder. 2018. Reddit: A gold mine for personality prediction. *Proceedings of the Second Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media (Association for Computational Linguistics)*.

Fei Liu, Julien Perez, and Scott Nowson. 2017. A language-independent and compositional model for personality trait recognition from short texts. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Association for Computational Linguistics)*.

Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E. Moghaddam, and Lyle Ungar. 2016. Analyzing personality through social media profile picture choice. *Association for the Advancement of Artificial Intelligence*.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*.

MBTI Database. https://www.mbtidatabase.com/.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter. *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (Association for Computational Linguistics)*.