Predicting MBTI Personalities based on Language Usage by Natural Language Processing Algorithms

Te-Yuan Chen, Yiyan Huang

The University of Texas at Austin

Abstract. Using machine learning to predict personalities is not only beneficial for customizing users' experience of interacting with products, but could further improve their experience of using artificiall intelligence products. This paper used natural language processing algorithms to predict MBTI personalities and extend the dataset scope to include Azure sentiment score and social media posts to testify prediction performance. It compared models of GRU, LSTM and RNN with different hyperparameters of node numbers, dropout regularization and epochs, and found the best prediction accuracy of 0.938. Meanwhile, it measured the influence of Azure sentiment score on prediction accuracy and the effectiveness of algorithm in predicting Tweets. Although results find the latter two efforts futile, the prediction results might be improved in the future by expanding dataset and incorporating more features.

1 Introduction

Motivation Personality categories are widely used for team building and division of tasks. Applications and products on social networks and team management tools also built their framework based on personality matching too. The application of personality prediction algorithm could also improve user experience with customized information presented, creating segmented markets and attaching characters to AI products with personality traits users are comfortable with. It could also address pain points such as those when users are asked to fulfill their own personalities, they expressed concerns of agnostics and dishonesty in fulfilling personality.

Studies showed high correlation between personality reflected and the language and text expressed by subjects. Two popular psychological personality assessments are the Big Five and Myers Briggs Type Indicator (MBTI). The Big Five measures personality with scores whilst MBTI defines 4 binary classes extending 4 scales, including Extraversion (E)-Introversion (I), Sensing (S)-Intuition (I), Perception (P) - Judging (J), Feeling (F)-Thinking (T). Previous studies also showed a higher feasibility of predicting MBTI in comparison with the Big Five. Using MBTI as the framework for measuring personality traits is more suitable for making predictions based on machine learning as well as future application in various products.

Literature Review Previous studies covered using Twitter and other social media to make MBTI predictions [1–4]. The algorithms they used to predict personality included Naive Bayes and Support Vector Machine(SVM)[5, 6]. However, we didn't see papers that use the RNN model to analyze language data and build a model. Meanwhile, we used the Kaggle open dataset with data collected through the Personality Cafe forum, whose MBTI labeling is more accurate because users labeled personalities by themselves. We also found previous studies only used traditional classification methods whilst the RNN model was rarely utilized.

To this end, we will use deep learning algorithms in natural language processing, i.e. RNN to predict the personality of people based on things users spoke or texted on the forum. We use a new dataset seldomly used in previous studies but with clear labels on MBTI. We expect to see a higher accuracy in this model because RNN is strong in analyzing time-series data, making classification predictions and widely used for natural language processing.

2 Related Work

Relations between Personality Categorization and Language Use. Studies on psychology and linguistics consolidated the correlation between usage of natural language and personality prediction. Previous studies by psychologists and linguists used content analysis of sample texts to match its consistency with the self-evaluation on personality modes of subjects. Robert and Douglas studied the relation between language people used for self-expression and the responsible personality traits of thinking-feeling scale of MBTI. They observed significant correlations between the content analysis of verbal samples and subjects' self-evaluation score on the thinking-feeling scale. Gender is also a factor they used in measuring subjects' MBTI scores [7]. Digman and Takemoto-Chock attempted to compare previous classical studies and extracted 5 factors in cognitive analysis consistent in previous studies [8]. Lee et al. used Korean language to study the connection between personality assessment results and language use. Studying Myers-Briggs (MBTI) and the 5-Factor Inventory allows them to find that personality was significantly related with linguistic factors [9]. Isbister and Nass's study showed that participants could accurately identify extroverted language as significantly more extroverted than the introverted language [10].

This work is developed based on theories on the connection between language and personality. Further, we will use algorithm to make predictions with reference to previous linguistic content analysis for predicting personality.

Personality assessment and social media. With the development of machine learning, the academic community started to explore the prediction of personality of authors across different media. Recently, machine learning for predicting personalities by using social media data is more popular. Celli and Lepri compared the accuracy of predicting 2 prominent psychological tests for personality

assessment and found out that MBTI with 4 binary classes defined and 16 personality types in combination. They found out that MBTI has better performance with algorithms trained than the Big Five [1]. Ahmad and Siddique studied the performance of gradient boosting algorithm trained with Kaggle MBTI dataset and spotted performance improvement[1]. Wang collected the latest 200 tweets from Twitter users with MBTI marks and studied the features of n-grams, Twitter POS tags, word vectors to make predictions [3].

Some previous work combine user profile on social media accounts and language for analysis, and most of these studies use Twitter data. This work used a new dataset with information from a personality forum.

Training NLP model to predict personality. Several pieces of work have used machine learning algorithms to learn from language and predict personality. In order to extract useful semantic features into models, there are some methods like LIWC, n-grams can help [11, 4]. After pre-processing, people will use algorithms like Naive Bayes and Support Vector Machine(SVM) to predict classification[5, 6]. Moreover, there are some models like Logistic Regression [4] and Gradient Boosting Model [12].

To be different from the above research, we will try to use the RNN/GRU/LSTM models for predicting personality with words. Furthermore, we will also try different feature extraction methods to get unique word features.

Application of Predicting Personality Algorithm with Social Media Data. Several studies explored the usage of personality predicting algorithm with using social media data. Golbeck et al. used people's personal profiles displayed online to predict their personalities and highlighted that the implications of this research include improving friends suggestions, revealing actual personalities of users to augment the popularity of platforms, online marketing, career planning and improving ways of information presented on social media to keep in line with personal traits of users [13]. There are also cases where the trained algorithm on predicting personality could be used in spam tracking, anti-social personality traits and education. Ezpeleta et al. used personality features to improve the prediction accuracy of spam messages [14]. Wald et al. predicted the Dark Triad traits, i.e. narcissism, Machiavellianism and psychopathy with social media data and explored the feasibility of identifying anti-social traits for psychological interruptions [5]. Tlili et al. studied how MBTI personalities influenced users' experience with computer-based learning tools [15].

This work is motivated by these work and hopes to further discuss the application of trained algorithm in predicting personality.

3 Methods

This research expects to 1) compare the performance on accuracy and precision of different algorithms with optimal hyperparameters in predicting personality,

2) train an algorithm based on the results in the previous step to make predictions of personalities, and 3) improve the prediction accuracy and precision than the baseline.

To realize this, we will first clean the data downloaded from Kaggle, including the label and the data. For the labels, the original dataset provided 16 categories of MBTI personality (such as INTP, ENTJ), which we will use one-hot encoding to code data that Tensorflow could recognize. For the data, the original dataset is sentences put up with by users, which includes a variety of punctuation, symbols, or distracting factors. We will remove or modify them to preprocess the data first to make them recognized by algorithm, then transform words into numbers and use the whole sentence as one feature as prevalently used in NLP for processing data. Afterward, we will split the dataset into training, validation, and test dataset (80/10/10).

Then we will train RNN, GRU and LSTM and find out the best hyperparameters to see which algorithm has the best performance in predicting the test dataset. We will run the other experiment by including the feature of "sentiment" extracted by Azure and explore whether the joining of this feature engenders better prediction results. We will also testify the application of this algorithm by inviting classmates to provide a sentence in their email or social media to see if the machine could predict their personalities correctly.

4 Experimental Design

Experiment1: Comparing the performance of different algorithms in predicting personalities We compared the performance of RNN, GRU, and LSTM in making predictions by using the training split for training and validation split for finding their optimal hyperparameters. We explored the optimal value of node number, regularization method of dropout, and epoch number for hyperparameters. The Dataset we used is the (MBTI) Myers-Briggs Personality Type Dataset posted by Mitchell J. on Kaggle, which contains the forum posts from users in different personalities. The baseline we made a comparison with is the performance of LSTM model (1 embedding layer + 1 LSTM layer + 1 DNN layer, hyperparameters include 64 nodes, relu as activation function, and adam as model optimizer) with an accuracy of prediction of 0.22 (https://www.kaggle.com/uplytics/tensorflo
2-0-lstm-model). The accuracy performance of algorithms on training and test splits is used as the evaluation metric.

Experiment2: Comparing the performance with features that include the sentiment score extracted from Azure with the performance of algorithms trained merely with the original dataset We used Azure cognitive service to extract the sentiment score of the sentences and included the score as a new feature to the original dataset. Using the optimal hyperparameters and algorithm from Experiment 1, we compared the prediction performance on two different datasets to see if sentiment could improve the accuracy. Datasets we used are the one same as above and one combined with Azure sentiment score. The baseline we

made a comparison with is the performance algorithms in the previous step. The accuracy performance of algorithms on training and test splits is used as the evaluation metric.

Experiment3: Predicting personality by inputting Twitter posts After the previous experiments and trained out the algorithm with the best performance, we randomly selected 16 people with 10 different personality categories by searching hashtag along with personality types, and collected his or her latest tweets to make a 500-word entry for each person. We expected to learn whether the trained algorithm above could make similar great predictions with the new Twitter dataset. Similar to above, the dataset we have been using is the Kaggle MBTI dataset and the baseline will be the accuracy of the algorithm trained with the best performance in testing and training.

5 Experimental Results

Experiment1 We testified the performance of different hyperparameters including epochs, nodes and dropout regularization with the validation split to compare prediction accuracies of different algorithms. To have a better view of the performance of regularization method, we also reported the performance of different combination on the test split. The results are shown as Fig. 1 (Figure 1). In general, we can observe that the prediction accuracy improves with increasing nodes. With the regularization of dropout, the algorithm performance on the test dataset is more satisfying. But when the epoch increases, the algorithm is trained overfitted and the accuracy drops. Comparing different algorithms, we found that both LSTM and RNN achieved the best performance in testing but LSTM has a more stable performance with or without regularization. In light of this, the combination of 256 nodes, 1 epoch, dropout regularization and LSTM delivered the best results both on training and testing. The best prediction accuracy is 0.938, higher than the baseline model LSTM model (1 embedding layer + 1 LSTM layer + 1 DNN layer, hyperparameters include 64 nodes, Relu as activation function, and Adam as model optimizer) with an accuracy of prediction of 0.2. We selected this combination to continue next experiments.

Due to the limit of GPU on computation, we only try out three algorithm with 3 hyperparameters. More attempts could be made in the future to include more algorithms and hyperparameters for making comparisons, such as SVM and other algorithms. For more prominent comparisons, change of algorithm structures and layers could also be considered.

Experiment2 We used Azure Cognitive Service to measure the sentiment score of each row of data, and combined the data into the original processed dataset to compose a new feature. Using the same parameters as Experiment 1, we found the accuracy scores of dataset with sentiment index are respectively 0.938 for validation and 0.934 for the test split, slight lower than that with original dataset. Instead of improving the prediction accuracy, the new feature of sentiment lowered the prediction performance on the test split. It indicates that

With Dropout Regularization

	LSTM			RNN			GRU		
	16	**128**	**256**	**16**	**128**	**256**	**16**	**128**	**256**
Epoch 1	0.938	0.938	0.938	0.938	0.937	0.938	0.938	0.938	0.938
Epoch 2	0.938	0.936	0.938	0.936	0.938	0.935	0.934	0.936	0.938
Epoch 3	0.932	0.927	0.937	0.935	0.936	0.938	0.927	0.934	0.937
Test	0.931	0.93	**0.938**	0.933	0.935	**0.938**	0.926	0.936	0.934

Without Dropout Regularization

	LSTM			RNN			GRU		
	16	**128**	**256**	**16**	**128**	**256**	**16**	**128**	**256**
Epoch 1	0.938	0.938	0.938	0.938	0.938	0.938	0.938	0.938	0.938
Epoch 2	0.936	0.934	0.938	0.938	0.937	0.938	0.938	0.936	0.91
Epoch 3	0.926	0.929	0.938	0.934	0.937	0.937	0.927	0.93	0.889
Test	0.928	0.93	**0.938**	0.933	0.937	0.935	0.926	0.935	0.891

Fig. 1. Prediction of accuracy with different algorithms and hyperparameters

the sentiment extracted from sentences does not have direct relationship with personality prediction.

In the future, we might consider to try out different features extracted by Azure to see if this would help further improve algorithm performance. Meanwhile, future studies could also explore more features extracted from the original data to see if the performance will be improved.

Experiment3 The predicted categories and actual labels are shown in Figure 2 (Figure 2). Statistically, the accuracy of prediction is 0. We unfortunately found none of the 16 labels are predicted correctly. But when we further dissected the prediction and actual label, we found in many cases, only one of the 4 binary categories is attached with a wrong label, such as ENTP vs. ENTJ, INFP vs. ENTP. We measured the accuracy of predicting values in the single category, and found the accuracy scores for E/I, N/S, T/F, and P/J are respectively 50.00%, 68.75%, 43.75%, 37.5%. It is likely to deduce that for the current algorithm, it is easiest and most likely accurate to predict the label of N/S (intuition vs. sensing), while most difficult to predict P/J (perceiving vs. judging). The reason for the low accuracy in predicting Tweets might be that social media posts are different from their posts on special psychological forums. It indicates that the purpose and channels of texts used by users might cause inaccuracy in personality prediction.

Future studies might explore different datasets and forms of text input to testify the performance of algorithms, such as transcripts from human interaction with Alexa and email texts.

Prediction	ENTP	ENTP	INFP	ISTP	ESFJ	ESFJ	INTP	ISFJ	ISFJ	INTP	ESTJ	ISFJ	ENTP	ENFP	ISTP	ISTP
Label	ENTJ	ENTJ	ENTP	INTJ	ESTP	ISTP	ESTJ	ISTP	ESFJ	ESFJ	ISFP	ISFP	INFP	INFP	INFP	INFP

Fig. 2. Predicted results and true labels of predicting Tweets.

6 Conclusions

This paper used natural language processing algorithms to attempt to predict people's personalities and extend the data scope to include Azure sentiment score and social media posts. When deciding on the algorithm to adopt and optimal hyperparameters, we found that compared to RNN and GRU, LSTM has a better performance in accuracy and stability on personality predicting as shown by the training and testing dataset. With optimizers of 256 nodes, dropout regularization and 1 epoch, the prediction accuracy could reach 0.938, higher than the baseline algorithm. Meanwhile, when we attempt to include a new feature, the sentiment score from Azure, the accuracy for training is not changed but that for testing is slightly dropped. It indicates the inefficiency of sentiment score in predicting personalities for the current algorithm. Although the prediction performance on Tweets is not very satisfying, this paper found different performance in predicting different categories. It could be a future direction of improving the algorithm performance. Also, the language and styles users adopted for different media platforms and channels varied, which led to the inaccurate prediction results. It would be interesting to further develop this research in the future by combining voice tones, visual movements and posts from different platforms to make more accurate predictions on peoples's personality.

References

1. Celli, F., Lepri, B.: Is big five better than mbti? a personality computing challenge using twitter data. In: CLiC-it. (2018)
2. Ahmad, N., Siddique, J.: Personality assessment using twitter tweets. Procedia computer science **112** (2017) 1964–1973
3. Wang, Y.: Understanding personality through social media. (2015)
4. Verhoeven, B., Daelemans, W., Plank, B.: Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In: Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari, Nicoletta [edit.]; et al. (2016) 1–6
5. Sumner, C., Byers, A., Boochever, R., Park, G.J.: Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In: 2012 11th International Conference on Machine Learning and Applications. Volume 2., IEEE (2012) 386–393
6. Cui, B., Qi, C.: Survey analysis of machine learning methods for natural language processing for mbti personality type prediction

7. Seegmiller, R.A., Epperson, D.L.: Distinguishing thinking-feeling preferences through the content analysis of natural language. Journal of personality assessment **51**(1) (1987) 42–52
8. Digman, J.M., Takemoto-Chock, N.K.: Factors in the natural language of personality: Re-analysis, comparison, and interpretation of six major studies. Multivariate behavioral research **16**(2) (1981) 149–170
9. Lee, C.H., Kim, K., Seo, Y.S., Chung, C.K.: The relations between personality and language use. The Journal of general psychology **134**(4) (2007) 405–413
10. Isbister, K., Nass, C.: Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. International journal of human-computer studies **53**(2) (2000) 251–267
11. Komisin, M.C., Guinn, C.I.: Identifying personality types using document classification methods. In: Twenty-Fifth International FLAIRS Conference. (2012)
12. Amirhosseini, M.H., Kazemian, H.: Machine learning approach to personality type prediction based on the myers–briggs type indicator®. Multimodal Technologies and Interaction **4**(1) (2020) 9
13. Golbeck, J., Robles, C., Edmondson, M., Turner, K.: Predicting personality from twitter. In: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, IEEE (2011) 149–156
14. Ezpeleta, E., Zurutuza, U., Hidalgo, J.M.G.: Short messages spam filtering using personality recognition. In: Proceedings of the 4th Spanish Conference on Information Retrieval. (2016) 1–7
15. Tlili, A., Essalmi, F., Jemni, M., Chen, N.S., et al.: Role of personality in computer based learning. Computers in Human Behavior **64** (2016) 805–813