Explainable MBTI Classification: Proposal

Kyungwook Lee
20193396
l1905kw@kaist.ac.kr

Jihyeon Lee
20203511
jihyeonlee@kaist.ac.kr

Minseok Choi
20208147
minseok.choi@kaist.ac.kr

Kyungduk Kim
Naver Corp.
kyungduk.kim@navercorp.com

Introduction

NLP has been applied to various areas of life, creating lots of new values. This study starts with the question: "Can one's writing disclose their personality?" We are interested in finding out if there is a strong relationship between one's use of language and their actual personality. Every human being has an instinctive desire to learn about their own personality along with personalities of people around them. If NLP technology can derive a person's personality from their writing, it can be used as an auxiliary tool to understand each individual, and going further, diversities in the world.

The theory of personality is very diverse depending on the perspective, such as *characteristic theory* and *behavioral approach*. Symbolically, the Myers-Briggs Type Indicator (MBTI) is one of the most well-known and widely used descriptors of personality type. MBTI is designed to be used more easily and usefully in everyday life based on C. G. Jung's psychological theory, describing the way people behave and interact with the world around them with four binary categories, as shown in Table 1.

Table 1: MBTI categories

Char	Metric
Energy	Extrovert/Introvert
Information	Sensing/iNtuition
Decision	Thinking/Feeling
Lifestyle	Judging/Perceiving

Four categories make a total of 16 types. The four categories combine to create synergy, creating different tendencies. For example, ESTP and ESTJ differ only in their last personality, but when combined, their individual tendencies vary significantly. Therefore, it was not easy to place 16 personalities in proper positions in coordinates, and it would be more socially meaningful if we extract representative features and relationships by applying NLP to this domain.

In a recently conducted study, Cui and Qi [1] applied numerous machine learning methods on the MBTI personality type prediction. Softmax classifier, Naive Bayes, SVM, and LSTM-based models were adopted, and the LSTM network brought the best performance of 38% test accuracy. In addition to previous work, we focus on explainable aspects and improving model performance.

Our two main goals are (1) to predict one's MBTI personality type and (2) to analyze more than what existing MBTI means. First, we classify the personalities of the people into 16 MBTI types from given texts. A number of experiments have been conducted to detect patterns in people's behavior and have demonstrated that there are different uses of language within each MBTI category. However, they showed somewhat unsatisfactory results in terms of 16 MBTI types classification, which means that they still have difficulty generalizing relationships between language usage and personality. Therefore, we develop a better classification model to understand human behavior. Second, we find

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

words or sentences which have the greatest impact on the output of our model. Conventional deep learning models have been considered as black box models because it is difficult to understand which features have had an effect on the results, making further analysis challenging. Thus, by developing an interpretable MBTI type classifier, we find a preferred word or sentence composition for a particular personality. It is expected that by finding relationships between words and personalities, a simpler and more approximate method of MBTI type classification, which is easier than the current MBTI test, can be suggested.

We explore various neural network models to predict people's MBTI types with their short written text. We formulated the task as a text classification problem having 16 MBTI types as target classes. We use a convolutional neural network for the baseline of the work [2, 3]. Then, we implement RNN-based methods such as LSTM networks to compare the performance. To find out whether there is a bottleneck problem of RNN models, we also test a self-attention-based encoder for representing the input text.

We also investigate our method to visualize how the model predicts people's MBTI types on the model's inference. Attention mechanism usually allows the model to focus on the relevant expressions of the input text as needed [4]. This makes the attention used as a tool for interpretation of the decision process of our model. We adopt the attention mechanism to show which part of the text plays an important role in predicting the type.

We employ the MBTI dataset available from Kaggle, an open-source data platform. The dataset contains texts written from individuals having each type of personality, carefully collected through the Personality Cafe forum. Specifically, the dataset is comprised of over 8,600 rows of data in which each row represents each person's four-letter MBTI code, as well as the last 50 things that they have posted on the forum. Additionally, we obtained speeches and writings from prominent historical figures of each different MBTI type. They may have a different personality type in reality; nevertheless, they serve as good representative examples of how each type speaks or writes based on one's personality.

References

[1] Cui, Brandon & Calvin Qi. Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction. Stanford CS229 final project

[2] Majumder, Navonil, et al. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems* 32.2 (2017): 74-79.

[3] Yoon Kim, Convolutional Neural Networks for Sentence Classification, In *EMNLP*, 2014

[4] Sofia Serrano, Noah A. Smith, Is Attention Interpretable?, In *ACL*, 2019.