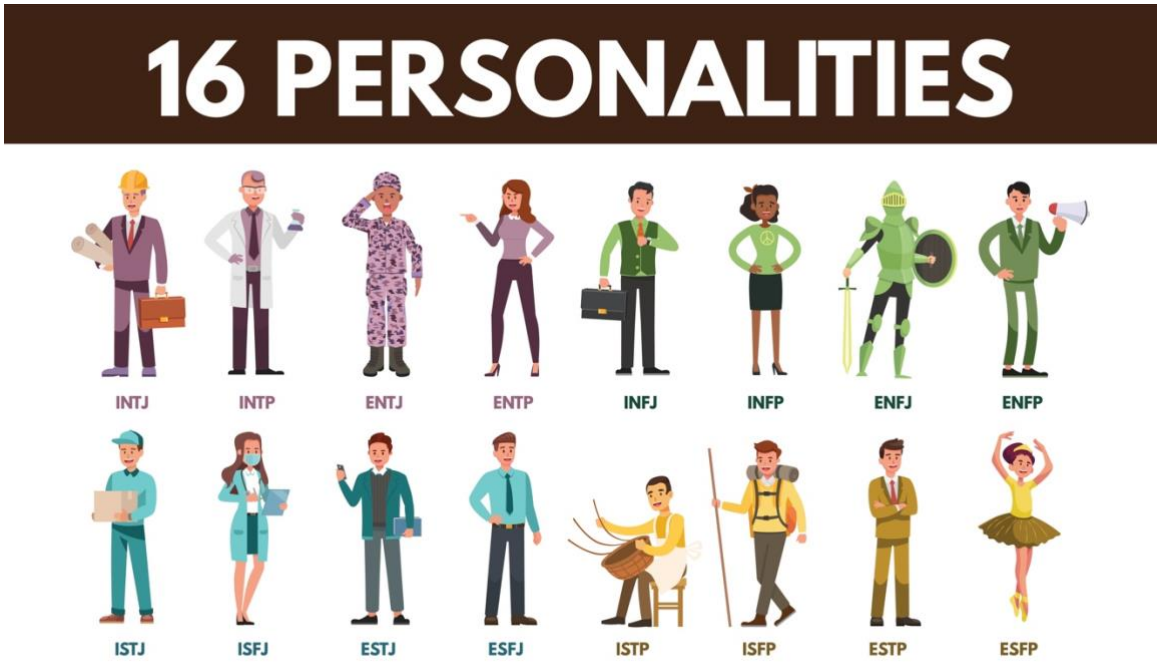
Using SHAP Explanations

Report On MBTI 500 Dataset

Omar Ahmed



Last Week's Discussion

In our previous discussion, we reviewed the outcomes of training the MBTI 500 dataset [~106 k instances] with various pre-trained transformer backbones for the Multi-Label Classifier.

- It was observed that all the different models used were all able to achieve high accuracies (94% and above). Such results raised suspension on the fidelity of the attained models.
- During the discussion and manual analysis of random samples from the dataset, keywords** referring to personality types such as INTJ, ESTP, etc., were identified. This prompted further investigation.

**Keywords: refer to the personality traits acronyms such as INTJ, ENTP, etc.

Training Another Model

After Identifying the keywords and finding out their high contribution to the dataset (fig.1). The idea was to train another model with the same data, but this time the keywords will be removed from the datasets before training the model.

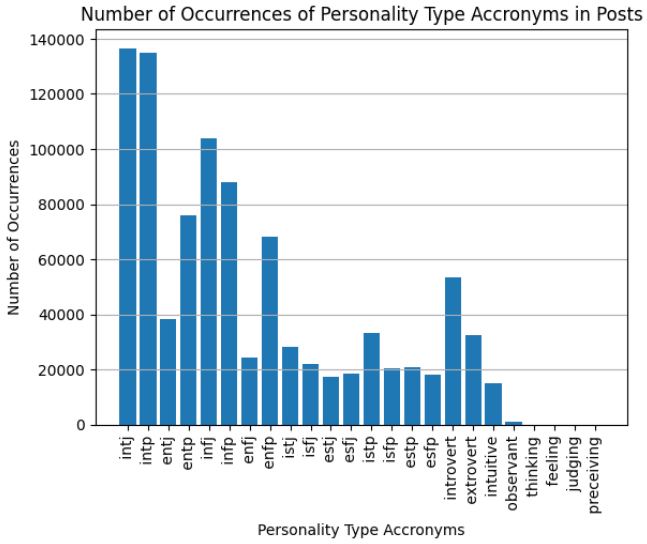


Figure 1

Training Results

After training the model on the updated dataset, its accuracy decreased to 87%. It is important to note that training was not the primary focus, so no hyperparameter fine-tuning was applied, and only 2 epochs were used for training. Instead, the aim was to focus on using explainability to investigate the effect of these keywords' presence in the training dataset on the model's predictions.

Explainability and SHAP values

After being able to apply SHAP Additive explanations to the multi-label classifier, it seemed reasonable to test and compare the differences in values assigned to the input based on the two models' predictions. Additionally, exploring how removing the personality keywords would shift the model's attention in the right direction.

The Experiment included picking a sample of 500 posts from the dataset and calculate the SHAP values for each of the 4 MBTI axes (IE, NS, TF, JP). The sample dataset had strong presence for the keywords (fig.2).

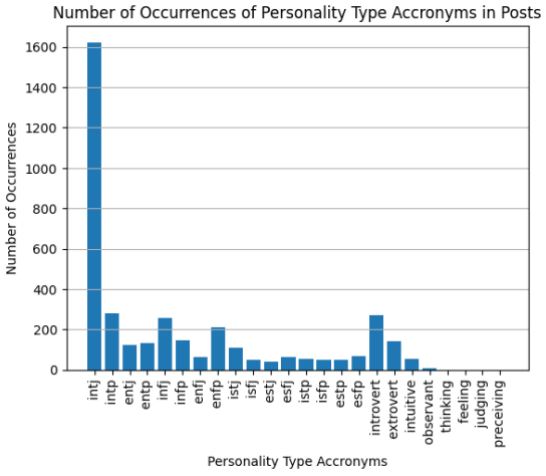
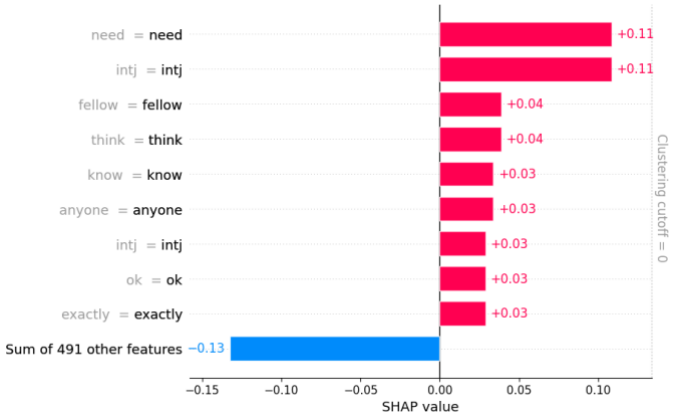


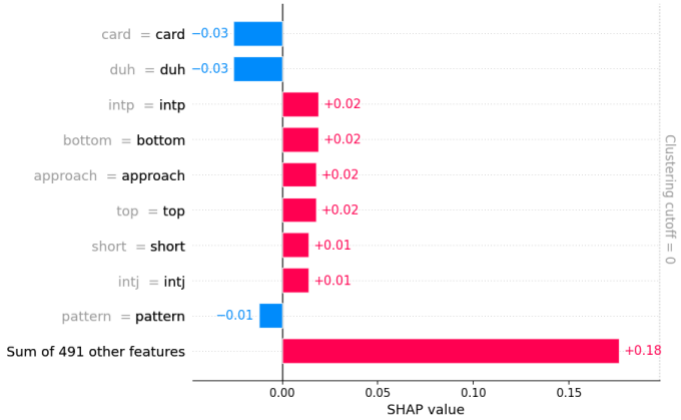
Figure 2

The Explanations results were as follows:

- When examining the SHAP values for random instances, the model trained on the raw dataset showed high relevance scores on different axes for the keywords as expected.

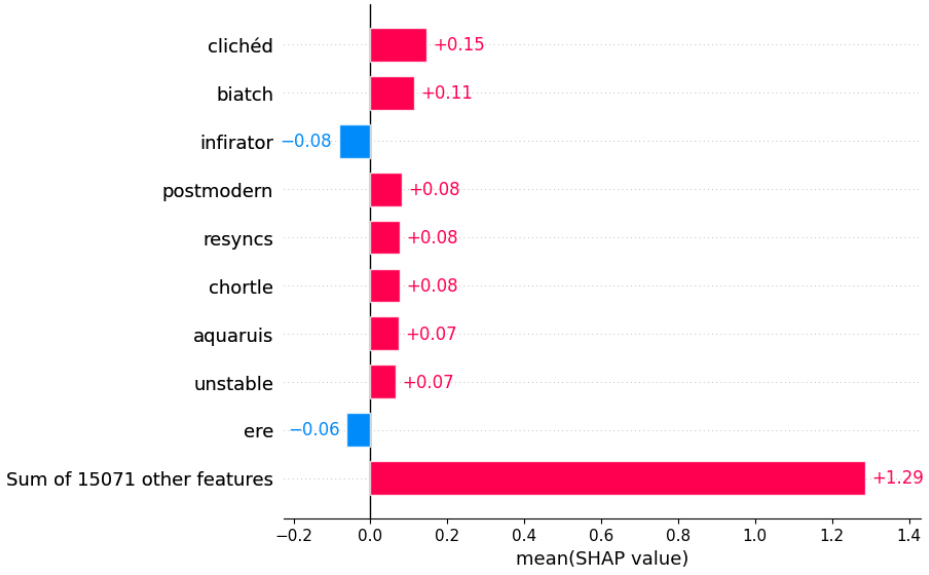


SHAP values IE axis, OUT: Introvert



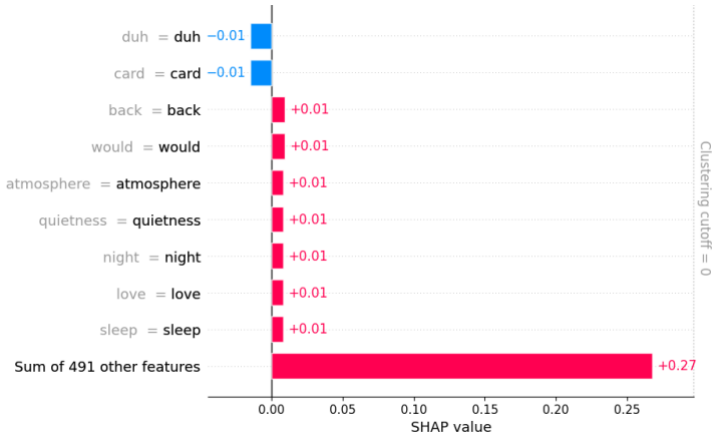
SHAP values JP axis, OUT: Judging

- However, Taking the average for the 500 samples shows absence of such keywords in the top relevant features.

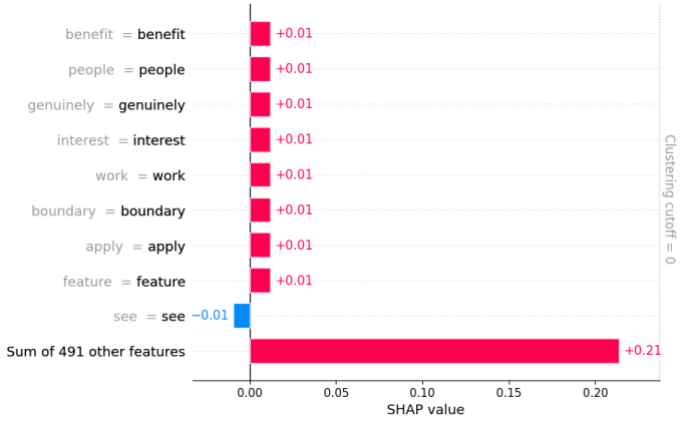


SHAP values mean on JP axis.

- Examining the same random samples for the model trained on the dataset with the words removed shows a shift for more plausible keywords.

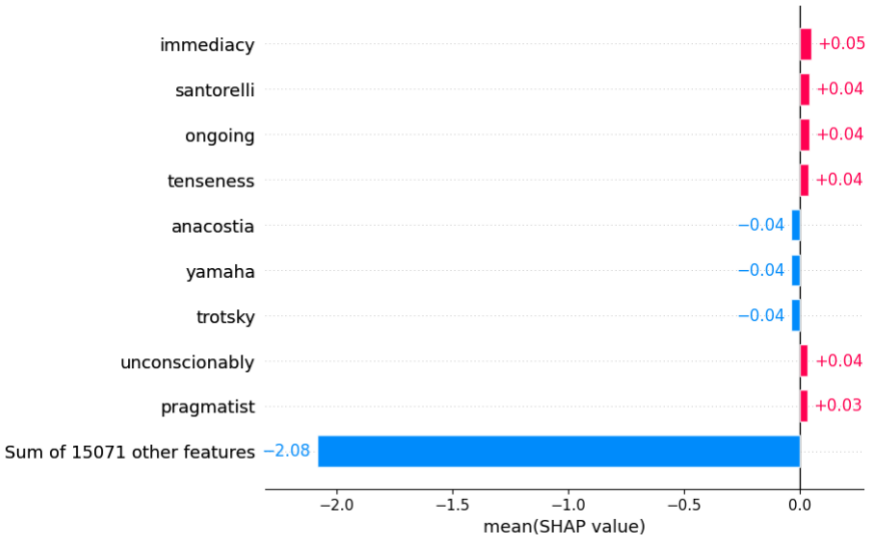


SHAP values IE axis, OUT: Introvert



SHAP values JP axis, OUT: Judging

- Although the average in the first model was not focusing on personality traits acronyms, the model with no keywords was able to focus on more plausible keywords.



SHAP values mean on JP axis.

Conclusion

In conclusion, exploring the MBTI 500 dataset revealed the significant impact of keywords on the model's predictions. Although the removal of these keywords resulted in a slight decrease in accuracy, it enabled the model to focus on more valuable features. This was further investigated through the calculation of SHAP values.

Limitations

A small test sample may not be representative, especially when it comes to the mean. Due to computational limitations, it takes 4 hours to produce SHAP values for 500 posts on the 4 MBTI axes.

References

- Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.
- SHAP documentation https://shap.readthedocs.io/en/latest/index.html
- MBTI 500 dataset https://www.kaggle.com/datasets/zeyadkhalid/mbti-personality-types-500-dataset/data
- Partition Explainer https://towardsdatascience.com/shaps-partition-explainer-for-language-models-ec2e7a6c1b77