Augmentation Techniques for Personality Type Classification Using Social Media Text

W.R.P.W.M.A.K.B. Dambawinna* and A.M.C. Malkanthi

Department of Physical Sciences & Technology, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka

* *avishkakavindud@gmail.com*

Social media platforms have evolved into a ubiquitous hub for individuals to convey their thoughts, emotions, and behaviours, allowing to generate insights into their personalities. However, the available social media datasets are not diverse enough as they often tend to over-represent certain groups of individuals such as young people. Further class distribution is disproportionate, potentially restricting the generalizability and accuracy of personality analysis. To address these challenges, this study suggests a method using text augmentation techniques and machine learning to expand the dataset and improve the effectiveness of the analysis using the Myers-Briggs Type Indicator (MBTI) dataset of 430000 posts belonging to 8600 individuals from personalitycafe forum. The dataset was split into 80% training and 20% testing sets and the training dataset was later augmented and fed into the models. All the models were evaluated using standard metrics such as accuracy, precision, and recall. Among the evaluated models, Linear Regression classifier outperformed the other three machine learning algorithms with an accuracy of 68.41%. The results are more uniformly distributed across the classes when compared with the other three machine learning algorithms Random Forest, Gradient Descent and XGBoost. Additionally, results showed that the text augmentation strategies employing BERT contextual word embeddings improved the model accuracy by 0.2%. A meagre improvement was observed due to the low quality of the dataset, and lack of contextual understanding in augmented data. Computational cost hindered the possibility of further improvement. Synonym-based augmentation showed poor performance due to a lack of contextual understanding, whereas BERT-based augmentation produced semantically and contextually relevant data, resulting in improved performance. For future work, self-training reinforced models and transfer learning need to be investigated to increase the model performance.

Key words: Deep Learning, MBTI, Personality Classification, Text Classification, Transfer Learning