Augmented Language Dataset for Enhanced Personality Profiling

Original Scientific Paper

Mohmad Azhar Teli*
Department of Computer Science;
University of Kashmir, Hazratbal Srinagar, Srinagar 190006, India
mohmadazhar.student@kashmiruniversity.net

Manzoor Ahmad Chachoo
Department of Computer Science;
University of Kashmir, Hazratbal Srinagar, Srinagar 190006, India
manzoor@kashmiruniversity.net

*Corresponding author

Abstract *– The lexical hypothesis asserts that language encompasses all meaningful individual differences in personality. Language is a vital tool for communication and self-expression, making it essential for understanding and assessing human personality. This paper investigates personality recognition from language use, emphasizing the significance of language in capturing and analyzing personality traits. A comprehensive literature review examines various approaches and techniques in personality recognition. We investigate the effectiveness of language use in predicting personality traits, employing multiple feature extraction and data augmentation techniques to enhance the accuracy and robustness of the personality recognition models. Our approach involves training a generative model, PersonaG, on the Essays dataset, subsequently using it to generate augmented data (AUG-Essays). We compare the performance of machine learning classifiers using LIWC, TF-IDF, Glove, and Word-Vec features on both Essays and AUG-Essays datasets. Our findings demonstrate significant improvements in predictive performance, offering valuable insights for applications in human resources, marketing, and beyond.*

Keywords: *Personality, Social Signal Processing, Natural Language Processing*

Received: July 16, 2024; Received in revised form: October 5, 2024; Accepted: October 7, 2024

1. INTRODUCTION

Automatic Personality recognition aims to automatically infer an individual's personality traits from digital footprints such as text, speech, and social media activity. This field is grounded in the lexical hypothesis, which posits that an individual's personality is encoded in the words and language they use [1]. Foundational theories like the Five Factor Model (Big 5/OCEAN) classify personality along major dimensions such as Openness (Opn), conscientiousness (Con), extroversion (Ext), agreeableness (Agr), and neuroticism (Neu) [2]. Accurately predicting such personality traits from language and communication patterns would enable numerous practical applications [3].

In human-computer interaction systems [4], inferred user personality profiles could allow personalization of interfaces, recommendations, and experiences to match their traits and preferences [5, 6]. Understanding customer personality derived from reviews, social posts, and surveys can inform targeted advertising and engagement strategies [7]. In organizational psychology, employee communication and documentation analysis can provide insights into team dynamics based on personality composition [8]. Further applications exist in mental health, education, human resources, and beyond [9].

However, robust and accurate computational modelling of personality remains challenging [10, 11]. Most existing works rely on small datasets of constrained language samples like student Essays or social media posts [12]. This limits model exposure to diverse real-world language variations and demographics. Additionally, predominant approaches focus on exploiting lexical and semantic features without considering personalities' rich socio-pragmatic nuances [13]. Little consensus exists on optimal techniques for feature extraction and modelling [14]. Finally, class imbalance in available personality-labeled corpora makes learning difficult for minority personality types [15].

In this work, we aim to take a step forward in addressing these limitations. Our contributions are three-fold: