Synthetic Data Generation with Large Language Models for Improved Depression Prediction

Andrea Kang
University of California, Los Angeles, Los Angeles, California, United States

Jun Yu Chen
UCLA, Los Angeles, California, United States

Zoe Lee-Youngzie
University of California, Los Angeles, Los Angeles, California, United States

Shuhao Fu
UCLA, LOS ANGELES, California, United States

Abstract

Automatic depression detection is gaining traction at the intersection of psychology and machine learning, but concerns over data privacy and scarcity persist. We propose a pipeline using Large Language Models (LLMs) to generate synthetic data that enhances depression prediction while addressing ethical concerns. Starting from recorded clinical transcripts, our chain-of-thought prompting involves two steps: the generation of a synopsis and sentiment analysis from the original transcript, and the generation of synthetic data based on these summaries and a new depression score. The resulting synthetic data not only performs well in terms of utility and fidelity, but also balances the severity distribution in training datasets, improving prediction of depression intensity. Our method offers a practical solution to augmenting limited, sensitive data while preserving statistical integrity. This framework provides a robust framework for advancing mental health research and applications without compromising patient privacy.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*