PeerJ Computer Science - Reviewing Manuscript 114547v1

“Classification of psychiatry clinical notes by diagnosis: A deep learning and machine learning approach”

This study compares different Artificial Intelligence (AI) models, including traditional Machine Learning (ML) approaches and advanced Deep Learning (DL) architectures, for classifying clinical notes on two disorders: Anxiety and Adaptation Disorder. These two disorders show partly overlapping symptoms but differ in some key features: in Anxiety Disorder, anxious symptoms are the core of the clinical picture, whereas in Adjustment Disorder, symptoms, which may include anxiety, depression, or behavioral changes, are triggered by identifiable stressful events. The main objective of the research is to evaluate the ability of the different AI models to accurately classify the two disorders based on the patients’ clinical notes. To this end, various oversampling techniques were employed to balance the dataset and optimize the performance of the models through a careful tuning phase of the hyperparameters.

1. Basic Reporting

The theme is interesting, and the manuscript is well-written. The introduction provides a strong background on the topic, with appropriate references and a comprehensive literature review. Moreover, the explanation of the paper's aim is clear. The figures and their captions are informative and useful.

2. Experimental Design

The study articulates a well-defined and pertinent research question, effectively justifying its necessity by delineating existing lacunae within the domain of AI-assisted psychological and psychiatric diagnostics. The authors provide a meticulous account of the targeted diagnoses, the utilized dataset, and the data cleansing procedures, including the terms of exclusion of clinical notes with a character count below 600. The methodology is comprehensively delineated, thereby ensuring replicability. Moreover, the ML models, encompassing Random Forest, Support Vector Classifier (SVC), Decision Tree, and XGBoost, alongside the DL models, SciBERT and DistilBERT, are elucidated with clarity, facilitating accessibility for researchers across diverse disciplines. The description of the dataset acquisition, employing the Electronic Health Record (EHR) format, is thorough and addresses privacy considerations adeptly. Additionally, the inclusion of hardware specifications enhances the potential for replication.

3. Validity of the Findings

The study demonstrates robust methodological rigor through the meticulous presentation of results. The inclusion of hyperparameter optimization and the application of diverse oversampling techniques enhance the study's analytical precision. Furthermore, validating the Large Language Model (LLM)-derived results by expert review, coupled with a rigorous preprocessing pipeline, substantiates the credibility of diagnostic extraction from individual clinical notes. The characterization of the dataset, specifically the distribution of gender, age, and diagnoses, facilitates a thorough understanding of the sample utilized in model training. Additionally, the study's acknowledgment of underrepresented demographics, particularly young and elderly populations, reflects a nuanced awareness of potential biases. The utilization of various oversampling methodologies, including the absence of oversampling, random oversampling, and Synthetic Minority Over-sampling Technique (SMOTE),

in conjunction with sample stratification and the application of a comprehensive suite of performance metrics, ensures the robustness of the findings. Finally, the strategic tuning of hyperparameters and the meticulous evaluation of performance metrics contribute significantly to the overall validity and value of the results obtained.

4. General Comments

The manuscript is well-structured and exhibits commendable readability, complemented by an appropriate selection of references. To enhance replicability, it is suggested that further details be provided regarding the hyperparameter tuning process. Additionally, Figures 4 and 5 would benefit from improved image quality, as their current resolution is suboptimal compared to the rest of the manuscript. Specifically, adopting a bar graph instead of a scatterplot with an arguably inappropriate trend line is recommended for an effective visual representation of the data. The comparative analysis of data and the elucidation of hyperparameter impact are acknowledged as valuable contributions. For future directions and limitations, it is recommended to consider the inclusion of an investigation into additional models, potentially incorporating clinical notes describing subjects with similar symptomatology who did not receive the diagnoses in question. Based on clinical notes, this approach could facilitate a more nuanced classification between disorder and non-disorder states.

Other minor issues:

a) In Figure 1, unities of measurement should be indicated (characters).

Overall, the study is well-conducted and contributes meaningfully to AI-assisted psychiatric diagnosis. A more detailed ethical compliance section and clearer data availability statement would improve the manuscript's transparency. No major methodological flaws were found, but minor clarifications would enhance replicability and robustness.

5. Confidential Notes to the Editor

None