GMR-159

LLM enabled Synthetic dataset generation for Human-AI teaming Algorithm

Author: Sai Sanjay Potluri, spotlur4@students.kennesaw.edu

Advisor: Dr. Md Abdullah Al Hafiz Khan

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Abstract Introduction

This research explores using Large Language Models (LLMs) to generate synthetic datasets for Human-AI teaming algorithms, focusing on mental health assessments. We create a diverse dataset simulating human-AI collaboration scenarios in diagnostic processes. The synthetic data is labeled through an innovative approach involving two human annotators and three LLMs, using majority voting for consensus-based annotations. This dataset serves as a resource for training and evaluating Human-AI teaming algorithms, enabling exploration of collaboration dynamics between human expertise and AI in complex decision-making. Our approach addresses the scarcity of real-world data in Human-AI teaming scenarios and provides a controlled environment for algorithm development, potentially accelerating advancements in this field.

Methods

An assessor is created which takes all the individual classifications as inputs and gives out a final classification using majority voting technique.

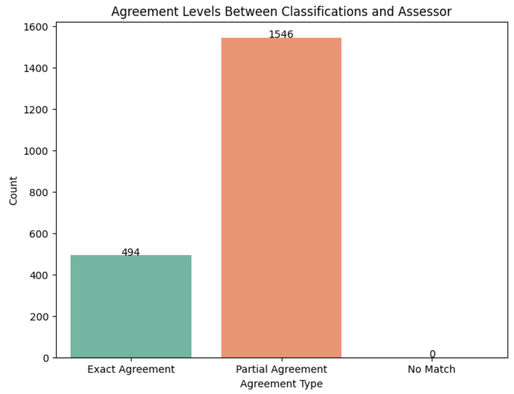


Fig.1: Agreement statistics between all classifications and assessor

Dataset description

1.Data Structure Overview:

Total Entries: 2040

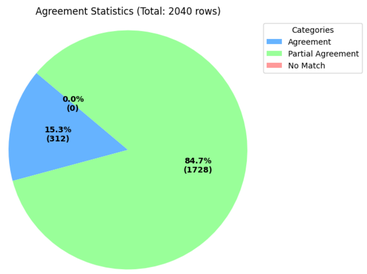
Columns: 12

Data Types: All columns are of object type (indicating text or categorical data).

2.Notable Columns:

description: 2040 entries, with 1994 unique descriptions. This suggests some repetition in the descriptions.

human_annotation: Categorical data with 8 unique classes, where "Anxiety" is the most frequent annotation.



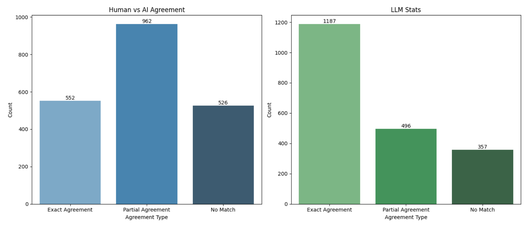


Fig.3: Comparison of Human Vs Ai and LLM comparisons statistics

phi_output, mistral_output, gemma_output: These columns store the AI model outputs and seem to have unique, detailed entries related to mental health classifications and analysis.

phi_classification, mistral_classification, gemma_classification: These contain the classification outcomes from different models, with "Anxiety" and "Depression" being common outputs.

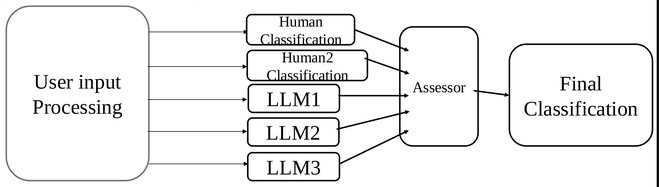
human2_classification: An additional human-annotated classification with 8 unique classes

assessor: A categorical feature with 18 unique values, likely indicating who or what assessed the data.

agreement_human_ai, agreement_llm: Indicate the level of agreement between human and AI or LLM classifications, with three possible values: "Exact Agreement," "Partial Agreement," and "No Match."

High level overview

- The user input from the description is the base text which is used for classifications.
- The user input is then annotated by both humans and LLMs after the input is passed as the prompt to the LLMs individually
- Then all the classifications are passed to an assessor which employs majority voting for determining the class which is agreed by most of the classifications for the given input.



Key Insights

- **Diversity in Classifications**: The dataset has a wide range of mental health classifications. The most frequent labels like "Anxiety" and "Depression" indicate a focus on these conditions.
- **Complex Output Fields**: phi_output, mistral_output, and gemma_output contain detailed, possibly unstructured text, likely summarizing AI or human analysis related to the mental health categories.
- **Agreement Levels**: There are columns dedicated to understanding how well human and AI classifications align, which can be crucial for assessing model performance.

Potential Next Steps for Analysis and Future Plans

- **Data Cleaning**: Address missing values in key columns (phi_output, phi_classification, and gemma_classification).
- **Text Analysis**: Use NLP techniques to analyze the description and AI outputs for patterns or to extract keywords and classifications.
- **Human Factors:** A human factors such as cognitive load, trust, etc., can be used on the labels after the majority voted labels
- **Model Performance Evaluation**: Investigate the agreement_human_ai and agreement_llm columns to understand how well AI models perform compared to human annotators.
- **Classification Distribution**: Explore the distribution of each classification category to see if there are any imbalances.
- Creating a neural network with a custom layer which is trained on the generated dataset for more precise classification
- By implementing more sophisticated methods, the research can be used to create a tool where both human and AI can agree on a middle ground in case of ambiguities.

Conclusions

Currently the research has 494 gold standard data where all the classifications and assessor are in absolute agreement i.e., both the human and all three LLMs identify and categorize it as the same classification.

Followed by 1546 silver standard data where there were one or more than one agreements among humans, LLMs and assessor

A positive outlook for the research is that there are no disagreements at least on the higher level where everything is compared with assessor classifications.

Literature Cited

- Gao, Y., Shen, Y., Gao, Y., Luo, X., Xiong, Y., Zhu, K. Q., & Gao, J. (2024). Generative AI for Synthetic Data Generation: Methods, Challenges and the Future. arXiv preprint arXiv:2403.04190.
- Confident AI. (2024, October 9). Using LLMs for Synthetic Data Generation: The Definitive Guide. https://www.confident-ai.com/blog/the-definitive-guide-to-synthetic-data-generation-using-llms
- Liu, Y., Ding, N., Chen, X., Wang, Y., Ding, Z., & Liu, Y. (2024). On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. arXiv preprint arXiv:2406.15126.
- IBM Research. (n.d.). Synthetic training data for LLMs. Retrieved November 4, 2024, from https://research.ibm.com/blog/LLM-generated-data
- JavaTpoint. (n.d.). Majority Voting Algorithm in Machine Learning. Retrieved November 4, 2024, from https://www.javatpoint.com/majority-voting-algorithm-in-machine-learning
- Raschka, S. (n.d.). EnsembleVoteClassifier: A majority voting classifier - mlxtend. Retrieved November 4, 2024, from https://rasbt.github.io/mlxtend/user_guide/classifier/EnsembleVoteClassifier

Contact Information

Connect with me on my LinkedIn using the QR code