MentalChat16K: A Survey on Conversational Mental Health AI

CMPE 255 - Data Mining Assignment

Student: Bala Anbalagan
Course: CMPE 255 - Data Mining Sec 47
Semester: Fall 2025
Topic: MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance

Paper Information

Attribute	Details
Title	MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance
Authors	Jing Xu, Tianming Wei, Bohan Hou, Patryk Orzechowski, et al.
Institution	University of Pennsylvania
arXiv Link	https://arxiv.org/abs/2503.13509
Dataset	HuggingFace - MentalChat16K
Official Code	GitHub - MentalChat16K

Overview

This repository contains my comprehensive analysis and presentation of the MentalChat16K paper, which introduces a benchmark dataset for developing AI-powered mental health conversational agents. The paper addresses a critical gap in mental health AI research by combining:

- **Real clinical data**: 6,338 QA pairs from 378 anonymized behavioral health intervention transcripts
- **Synthetic data**: 9,775 QA pairs generated using GPT-3.5 Turbo covering 33 mental health topics
- **Total**: 16,000+ question-answer pairs for training empathetic AI assistants

Why This Paper Matters for Data Mining

Data Mining Aspect	Application in Paper
Data Collection & Curation	Multi-source dataset creation (real + synthetic)
Data Preprocessing	Privacy-preserving paraphrasing using local LLMs
Data Quality	Manual filtering and de-identification pipelines
Feature Engineering	7 therapeutic evaluation metrics

Data Mining Aspect	Application in Paper
Model Evaluation	Multi-evaluator framework (GPT-4, Gemini, Human)
Benchmark Creation	Standardized evaluation for mental health AI

Repository Structure

```
MentalChat16K/
├── README.md                 # This file
├── slides/
│   └── MentalChat16K_Presentation.pdf    # Slide deck
├── images/
│   ├── architecture.png      # System architecture diagram
│   ├── dataset_composition.png
│   ├── evaluation_metrics.png
│   └── results_comparison.png
├── notebooks/
│   └── data_exploration.ipynb  # Optional: Dataset exploration
├── code/
│   └── evaluation_demo.py      # Optional: Evaluation demo
├── ARTICLE.md                # Medium article draft
└── VIDEO_LINK.md             # Link to video presentation
```

Key Contributions of the Paper

1. Dataset Creation

- First large-scale mental health dialogue dataset combining real and synthetic data
- 16,000+ QA pairs covering depression, anxiety, grief, relationships, and more
- Doubles the size of previous comparable datasets (Psych8K)

2. Privacy-Preserving Pipeline

- Uses local Mistral-7B for paraphrasing sensitive clinical transcripts
- Avoids uploading patient data to commercial APIs
- Manual de-identification of PII (names, addresses, financial info)

3. Novel Evaluation Framework

Seven mental health-specific metrics:

Metric	Description
Active Listening	Reflects and validates user concerns
Empathy & Validation	Shows understanding of emotional states

Metric	Description
Safety & Trustworthiness	Prioritizes user safety, suggests professional help
Open-mindedness	Non-judgmental, accepting of diverse perspectives
Clarity & Encouragement	Clear communication, positive reinforcement
Boundaries & Ethics	Maintains appropriate professional boundaries
Holistic Approach	Considers overall well-being

4. Comprehensive Benchmarking

- Fine-tuned 7 different 7B-parameter LLMs using QLoRA
- Compared performance with GPT-4, Gemini Pro, and human evaluators
- Statistical significance testing across all metrics

Methodology

Data Pipeline

```
┌─────────────────┐     ┌─────────────────┐
│  Clinical Trial │     │  GPT-3.5 Turbo  │
│   Transcripts   │     │   Generation    │
│   (378 sessions)│     │  (33 topics)    │
└────────┬────────┘     └────────┬────────┘
         │                       │
         ▼                       ▼
┌─────────────────┐     ┌─────────────────┐
│  Local Mistral  │     │  Airoboros      │
│  Paraphrasing   │     │  Framework      │
└────────┬────────┘     └────────┬────────┘
         │                       │
         ▼                       ▼
┌─────────────────┐     ┌─────────────────┐
│  6,338 QA Pairs │     │  9,775 QA Pairs │
│   (Interview)   │     │   (Synthetic)   │
└────────┬────────┘     └────────┬────────┘
         │                       │
         └───────────┬───────────┘
                     ▼
          ┌─────────────────┐
          │  MentalChat16K  │
          │  16,113 QA Pairs│
          └─────────────────┘
```

Fine-tuning Configuration

Parameter	Value
Method	QLoRA (Quantized Low-Rank Adaptation)
Models	LLaMA-2-7B, Mistral-7B, Vicuna-7B, Zephyr-7B
Hardware	NVIDIA A100 (80GB)
Training Configs	Synthetic only, Interview only, Combined

Results Summary

Key Findings

1. **Fine-tuned models significantly outperform base models** across all 7 metrics
2. **GPT-4 evaluator** favored synthetic data fine-tuning (alignment with GPT-3.5 patterns)
3. **Gemini Pro evaluator** valued real interview data, especially for safety metrics
4. **Human evaluators** consistently preferred fine-tuned models
5. **Combined training** did not always outperform individual approaches

Performance Comparison

Model Type	Avg Score (GPT-4)	Avg Score (Gemini)	Avg Score (Human)
Base Models	~6.5	~6.2	~5.8
Synthetic Fine-tuned	~8.2	~7.5	~7.8
Interview Fine-tuned	~7.8	~8.0	~7.6
Combined Fine-tuned	~8.0	~7.8	~7.7

Deliverables

1. Medium Article

Link: [Add your Medium article link here]

A comprehensive article covering:

- Introduction to mental health AI challenges
- Dataset creation methodology
- Evaluation framework explanation
- Key results and implications
- Personal analysis and future directions

2. Slide Presentation

SlideShare Link: [Add your SlideShare link here]

PDF Location: slides/MentalChat16K_Presentation.pdf

3. Video Presentation

Duration: 10-15 minutes

Link: [Add your video link here]

How to Use This Repository

Explore the Dataset

```python
from datasets import load_dataset

# Load MentalChat16K from HuggingFace
dataset = load_dataset("ShenLab/MentalChat16K")

# View sample
print(dataset['train'][0])
```

Run Evaluation Demo (Optional)

```bash
cd code
python evaluation_demo.py
```

References

1. Xu, J., Wei, T., Hou, B., et al. (2025). MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance. arXiv:2503.13509
2. Dettmers, T., et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs
3. Touvron, H., et al. (2023). LLaMA 2: Open Foundation and Fine-Tuned Chat Models
4. Chen, J., et al. (2023). ChatPsychiatrist: Evaluating LLMs for Mental Health Applications

---

License

Contact

Student: [Your Name]
Email: [Your SJSU Email]
LinkedIn: [Your LinkedIn Profile]

Last Updated: November 2024