Mental Health Signal Detector: Intent, Concern Classification, and Suggestion via RAG in Online Support Groups

Project Progress Report

Gaurav Zanpure, Shreyansh Kabra, Pragya Dhawan, Suyash Roy, Vanshika Wadhwa, Punith Basavaraj
University of Southern California
{gzanpure, kabras, pragyadh, suyashro, wadhwav, punithba}@usc.edu

Abstract

We developed a **Mental Health Signal Detector** that classifies posts in online support groups by their *intent* (e.g., Seeking Help, Venting, Giving Support) and *concern level* (Low/Medium/High). Building on our initial proposal, we have completed dataset curation, annotation, and baseline benchmarking. Early results from MiniLM and RoBERTa-Large models show strong generalization in concern detection ($\sim$0.69–0.78 F1). The next stage integrates a retrieval-augmented generation (RAG) component to produce concise, empathetic, and safe responses. This report summarizes our progress to date, dataset workflow, experimental results, and upcoming goals toward RAG-based generation.

Keywords: Mental Health, Intent Classification, Concern Detection, RAG, Benchmarking, Empathy Generation

1 Motivation

Online Reddit support communities such as `r/depression`, `r/anxiety`, and `r/suicidewatch` enable individuals to share emotional struggles and seek support. However, limited moderator bandwidth often leads to overlooked high-risk posts. Our objective is to design an interpretable, ethically responsible NLP system that automatically classifies such posts and generates empathetic, context-aware responses. The system aims to:

- Detect **intent(s)** (e.g. Seeking Help, Critical Risk, Cause of Distress).
- Detect **concern level** (Low, Medium, High).
- Generate empathetic, retrieval-grounded suggestions via **RAG**.

The system ultimately serves as a triage and moderation assistant, emphasizing interpretability, transparency, and psychological safety.

2 Literature Review

Research in computational mental-health NLP has transitioned from binary sentiment detection to nuanced, multi-intent understanding. The CLPsych Shared Tasks (2) and the eRisk Challenge (3) established foundational benchmarks for early risk prediction from Reddit and Twitter posts. Transformer-based architectures such as BERT and RoBERTa quickly surpassed traditional TF–IDF and SVM baselines, though domain adaptation remained critical.

Liyanage et al. (1) introduced a prompt-based augmentation framework for classifying *Wellness Dimensions (WD)*—Physical, Intellectual/Vocational, Social, and Spiritual/Emotional. Using GPT-3.5-turbo and GPT-3 variants (`text-curie-001`, `text-davinci-003`), their approach improved F1 by **13.1%** and MCC by **15.9%** compared to traditional augmentation techniques such as EDA and Back Translation.

Domain-specific encoders such as **MentalBERT** and **MentalRoBERTa** (4) further enhanced emotion and risk-detection performance through large-scale pretraining on Reddit corpora. Recent works emphasize ethical and safe NLP design for clinical applications (Valdez et al., 2023; Harrigian et al., 2022), while retrieval-augmented generation (RAG) architectures (5) combine factual grounding with empathetic response synthesis.

Our project extends these directions with post-level classification of **intent(s) and concern**, LoRA-based parameter efficiency, and retrieval-augmented empathy generation—unifying explainability and ethical AI design.

3 Dataset Collection and Annotation

We curated a multi-source corpus (2022–2025) comprising mental-health and support-focused Reddit posts.

- **Reddit Mental Health Dataset (Zenodo, 2020)** – foundational corpus from depression and anxiety subreddits.
- **Sentiment Analysis for Mental Health (Kaggle, 2023)** – labeled posts for Depression, Anxiety, Bipolar, Normal.
- **Reddit Depression Cleaned (HuggingFace, 2024)** – cleaned, deduplicated dataset with metadata.
- **Wellness Dimensions (Garg et al., 2023)** – contextual dataset for augmentation.
- **Mental Disorders Identification Reddit NLP (Kaggle, 2022)** – unlabeled corpus used for domain enrichment.

After normalization (lowercasing, URL/emoji removal), we consolidated **6,000 posts** containing titles and text bodies.

3.1 Annotation Process

A subset of **500 posts** was manually annotated using the **Label Studio** interface. Three annotators labeled each post for intent(s) and concern level. The annotations were cross-verified and finalized through discussion to ensure consistency and reduce subjective variation across annotators.

- **Intent(s):** Multi-label (2–3 tags/post) across nine categories: Cause of Distress, Miscellaneous, Mental Distress, Positive Coping, Critical Risk, Progress Update, Seeking

Help, Maladaptive Coping and Mood Tracking.

- **Concern Level:** Single label (Low, Medium, High) indicating distress and urgency.

Sensitive details were anonymized before annotation. A weakly supervised labeling step expanded to ∼8k posts using classifier confidence thresholds (≥0.85). Stratified splits of **70% train**, **15% validation**, and **15% test** were used to ensure each intent and concern level was proportionally represented. Concern-balanced stratification ensures the model encounters equal exposure to low, medium, and high-distress samples—preventing bias toward low-risk posts. The complete dataset preprocessing scripts, model configurations, and experimental pipeline are available at: **https://github.com/Gaurav-Zanpure/group-23-mh-signals**

4 Methods

A modular pipeline was developed using the Hugging Face Transformers library. Each classifier (Intent(s), Concern) uses a sigmoid-activated binary cross-entropy objective. Separate models are trained per task to preserve interpretability and reduce inter-label interference.

4.1 Model Hierarchy

- **Baseline: MiniLM + Logistic Regression** – mean-pooled `all-MiniLM-L6-v2` embeddings (384-dim) trained with a Logistic Regression head; transparent and fast.
- **Medium: DistilRoBERTa + LoRA** – parameter-efficient fine-tuning (rank = 64, α = 128) for mid-sized GPU setups.
- **Strong: RoBERTa-Large + LoRA** – rank = 16, α = 32, dropout = 0.1, achieving richer contextualization while reducing GPU usage by ∼60%.

All models are trained using AdamW Optemizer (lr = 1e-4, batch = 8, epochs = 8) with Macro-/Micro-F1 as the main metrics. Evaluation emphasizes calibration and interpretability.

5 Experiments and Results

Model	Intent F1	Concern F1
MiniLM + Logistic Regression	0.58	0.69
DistilRoBERTa + LoRA	0.64	0.78
RoBERTa-Large + LoRA	0.70	0.76

Table 1: Macro-F1 scores for intent and concern classification benchmarks.

5.1 Observations

- **Concern Classification:** DistilRoBERTa + LoRA achieves the highest concern-level performance (0.78 F1), showing excellent generalization across Low, Medium, and High distress categories. Most confusion occurs between Medium and High levels, where lexical tone and emotional intensity overlap.
- **Intent Classification:** With nine intent labels, multi-label prediction remains challenging. Semantic overlap among "Cause of Distress," "Maladaptive Coping," and "Mental Distress" causes boundary ambiguity, especially in shorter posts. RoBERTa-Large + LoRA provides the strongest contextual discrimination (0.70 Macro-F1), outperforming smaller models.
- **MiniLM Baseline:** Despite its simplicity, MiniLM + Logistic Regression achieves stable feature separability (0.58 Intent F1, 0.69 Concern F1), serving as a transparent, interpretable baseline.
- **LoRA Efficiency:** Parameter-efficient fine-tuning with LoRA enables both medium and large models to converge rapidly while reducing GPU memory usage by $\approx 60\%$, confirming scalability for resource-constrained training.
- **Label Distribution Impact:** The stratified 70/15/15 split by concern level ensures balanced representation across distress tiers, preventing bias toward low-severity samples and improving overall model robustness.

6 Next Steps

Before the final report, we will focus on the generation phase to extend classification outputs into actionable, empathetic text suggestions. Planned milestones include:

1. Collect and structure a corpus of verified coping and empathy response templates to power the RAG module.
2. Build a retrieval-augmented generation (RAG) pipeline that conditions suggestions on detected intent and concern.
3. Fine-tune `Flan-T5-small` on the curated corpus for concise, safe generation.
4. Conduct qualitative human evaluation for empathy, safety, and contextual alignment.

7 Progress and Future Contribution

So far, we have:

- Collected, cleaned, and merged a unified corpus of **6,000 Reddit posts** focused on mental-health discourse.
- Completed manual and semi-supervised labeling across both intent (9 tags) and concern (3 levels), ensuring balanced representation.
- Benchmarked three model families — **MiniLM**, **DistilRoBERTa**, and **RoBERTa-Large** — achieving strong results with RoBERTa-Large + LoRA (**0.70 Intent F1**, **0.76 Concern F1**).
- Established a reproducible, modular pipeline for intent and concern classification, supporting easy future integration of the **RAG-based empathetic suggestion** component (to be developed next phase).

Before the final report, we will focus exclusively on **building and evaluating the RAG-based empathy generation module**. This includes collecting safe coping-response data, developing retriever logic, and integrating Flan-T5-small for context-aware generation. These efforts will complete the full end-to-end empathetic moderation system.

8 References

[1] C. Liyanage, M. Garg, V. Mago, and S. Sohn, "Augmenting Reddit Posts to Determine Wellness Dimensions impacting Mental Health," in Proc. Conf. Assoc. Comput. Linguist. BioNLP Workshop, Jul. 2023, pp. 306–312. DOI: 10.18653/v1/2023.bionlp-1.27.

[2] P. Resnik et al., "The CLPsych Shared Task Series: Applying NLP to Mental Health," CLPsych Workshop, 2015–2021.

[3] D. Losada et al., "eRisk 2018: Early Risk Prediction on the Internet," CLEF Labs, 2018.

[4] S. Ji et al., "MentalBERT and DistilMentalBERT: Domain-specific Pretrained Transformers for Mental Health Texts," EMNLP, 2022.

[5] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.

[6] D. Valdez et al., "Ethical NLP for Mental Health: Balancing Privacy, Safety, and Interpretability," ACL, 2023.