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

Project Proposal

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 propose a real-time system to analyze short posts in online support groups, classifying them by *intent* (Seeking Help, Venting, Giving Support) and *concern level* (Low/Medium/High), and highlight key phrases that drive predictions. Additionally, a lightweight retrieval-augmented generation (RAG) module produces brief, safe, contextually relevant suggestions. We will benchmark three models of increasing complexity and evaluate them on accuracy, calibration, latency, and suggestion relevance, providing a practical end-to-end pipeline for mental health support.

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

1 Motivation

In large online support communities, urgent or high-risk posts are often overlooked due to high volume and limited moderator attention. This poses a risk to timely interventions, particularly when members express severe concerns. To address this challenge, we propose a calibrated pipeline that predicts the *intent* and *concern level* of a support group post, highlights key aspects, and generates a brief, safe suggestion. By incorporating RAG and appending a crisis disclaimer, the system can assist moderators in prioritizing responses and ensuring that critical posts receive timely support.

2 Literature Review

According to [1], the authors propose a prompt-based generative NLP approach to augment Reddit posts for classifying *Wellness Dimensions (WD)* that impact mental health. The dataset, collected from `r/depression` and `r/suicidewatch`, contained four imbalanced classes—Physical Aspect (PA), Intellectual and Vocational Aspect (IVA), Social Aspect (SA), and Spiritual/Emotional Aspect (SEA). To address this, they experimented with ChatGPT models (gpt-3.5-turbo, gpt-3.5-turbo-0301) and other GPT-3 variants (`text-curie-001`, `text-davinci-003`) to generate synthetic but semantically consistent training data. The augmentation process involved carefully designed prompts that produced both synthetic text and short explanations, enabling balanced class distributions. Their approach was compared with conventional augmentation techniques such as Easy Data Augmentation (EDA) and Back Translation (BT). The augmented data was then used to fine-tune a BERT classifier, where ChatGPT-augmented samples outperformed baselines, showing F-score and Matthews Correlation Coefficient (MCC) improvements of up to 13.11% and 15.95% respectively. This study highlights that generative large language models provide superior diversity and contextual alignment in text augmentation for mental health NLP tasks compared to traditional methods.

3 Data

We will use public Reddit-style datasets (e.g., Stress/Anxiety, SuicideWatch, Depression) and add manual multi-label annotations for intent (2–3 tags per post). Concern levels are annotated as Low, Medium, or High. A fixed train/validation/test split will be maintained.

Datasets: Reddit Depression Dataset, Reddit Anxiety Dataset, Reddit Public Comments Dataset, Sentiment Analysis for Mental Health, Mental Disorders Identification Reddit NLP etc.
Labels: Multi-label intent annotations (2–3 tags per post).
Concern Level: Annotated as Low, Medium, or High.
Splits: A fixed train/validation/test split.

4 Methods

- Baseline: MiniLM embeddings + Logistic Regression
- Medium: DistilRoBERTa + LoRA fine-tuning
- Strong: RoBERTa-base + LoRA

Suggestions are retrieved from a curated corpus of safe coping snippets based on the post's intent and concern, with a visible safety disclaimer.

5 Experiments

All models will be trained on the same splits with consistent preprocessing. Metrics: Macro-/micro-F1 and PR-AUC for intent; Macro-F1 and PR-AUC for concern; calibration error, latency; suggestion relevance is human-judged. We will also measure suggestion relevance (human-judged) and ensure retrieval faithfulness. Error analysis will focus on common failure modes.

6 Plan

Week 1: Finalize labels, splits, baseline; Weeks 2–3: Train DistilRoBERTa and RoBERTa; Week 4: Add calibration and highlights; Week 5: Implement suggestion retriever and UX mockup; Week 6: Human evaluation and report.

7 Contribution

Our project provides a benchmark across three modeling strategies for intent and concern detection, demonstrates a safe retrieval-based suggestion step, and integrates explainability and calibration into a practical end-to-end pipeline.

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.