Emotion-Aware Retrieval-Augmented Generation for Mental Health Support from Reddit Posts

Sai Siddu Vardhan Reddy Annadi, Vinay Kumar Reddy Punuru, Umair Moeen Tajmohammed

Abstract:

Supporting individuals facing mental health challenges through AI systems requires a delicate balance between empathy, contextual awareness, and factual grounding. We present an Emotion-Aware Retrieval-Augmented Generation (RAG) system that addresses these needs by integrating emotion classification with semantic retrieval and lightweight generative models. Our pipeline classifies user queries into one of 28 emotion categories using a fine-tuned RoBERTa-GoEmotions model, retrieves top-k emotionally and semantically similar Reddit posts using FAISS and sentence embeddings, and generates grounded, empathetic responses using large language models (Phi-3 Mini, Falcon-7B-Instruct). Evaluation across multiple metrics, including BLEU, ROUGE, and Precision@k, shows that emotion-aware retrieval significantly improves the coherence, safety, and emotional alignment of generated responses. Our findings demonstrate the importance of combining retrieval and emotional context for sensitive domains like mental health.

Introduction:

Large language models (LLMs) have revolutionized natural language generation, exhibiting impressive fluency and contextual understanding. However, their deployment in emotionally sensitive applications, such as mental health support, is fraught with challenges. These include hallucinations, emotionally mismatched responses, and the risk of unsafe advice. In this context, ensuring empathy, safety, and factual alignment becomes critical.

Retrieval-Augmented Generation (RAG) offers a promising solution by grounding model responses in real, human-generated content. When paired with emotion detection, this approach can tailor responses not only to the query's content but also to the user's emotional state. In this work, we present a novel Emotion-Aware RAG system that classifies the emotional tone of a user query and retrieves semantically and emotionally aligned Reddit posts to construct more personalized and contextually grounded responses.

We leverage RoBERTa-GoEmotions for emotion classification, all-MiniLM-L6-v2 for sentence embedding, FAISS for efficient nearest-neighbor retrieval, and Phi-3 Mini and Falcon-7B-Instruct as our generative backends. Our experimental evaluation highlights that emotion-aware conditioning, combined with retrieval, significantly enhances the relevance, safety, and helpfulness of responses in mental health scenarios.

System Overview:

Our system is designed around a modular Emotion-Aware Retrieval-Augmented Generation (RAG) pipeline that supports mental health queries with empathetic, grounded, and emotion-aligned responses. The architecture integrates emotion detection, semantic retrieval, and language generation components to create safe and contextually accurate outputs.

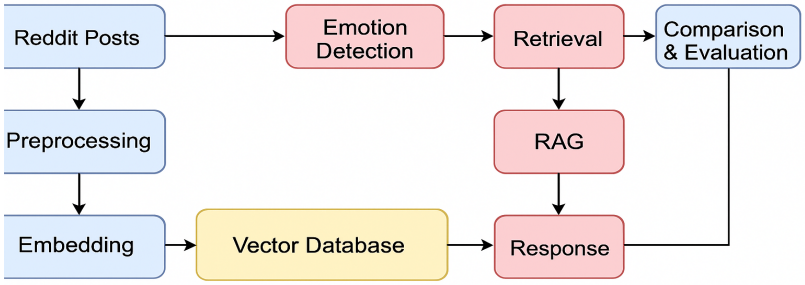

Figure 1: RAG System architecture

1. Dataset:

We utilize the Kaggle Reddit Mental Health Support Dataset, which comprises thousands of user-generated posts from mental health-related subreddits. To ensure high-quality inputs, we perform extensive preprocessing. This includes converting all text to lowercase, removing line breaks, and filtering posts to retain only those explicitly related to mental health discussions. We also eliminate noisy elements such as URLS, emojis, and special characters to enhance the clarity and semantic quality of the data.

2. Emotion Detection:

For emotion classification, we adopt the RoBERTa-base language model fine-tuned on the GoEmotions dataset (Demszky et al., 2020). This model is capable of identifying 28 nuanced emotion categories, such as *sadness*, *optimism*, *anger*, *gratitude*, and *anxiety*. Upon receiving a user query, the model assigns an emotion label that guides the subsequent retrieval and generation stages to ensure emotionally aligned outputs.

3. Embedding and Retrieval:

We apply the all-MiniLM-L6-v2 model from the Sentence Transformers library to encode both the user queries and the Reddit corpus into dense semantic embeddings. These embeddings are indexed using FAISS (Facebook AI Similarity Search), enabling efficient approximate nearest neighbor search. At runtime, given a user query and its detected emotion label, the system retrieves the top-k ($k = 3$) semantically and emotionally similar posts from the index.

4. Response Generation:

The response generation component of our system integrates two large language models (LLMs), carefully selected to strike a balance between fluency, empathy, and factual grounding. The first model, Phi-3 Mini (Microsoft, 2024), is a lightweight, safety-aligned LLM optimized for producing grounded and emotionally supportive responses. It emphasizes empathy and is designed to minimize hallucinations, making it suitable for sensitive mental health interactions. The second model, Falcon-7B-Instruct (developed by the Technology Innovation Institute, UAE), is a more expressive and fluent model capable of generating rich, detailed responses. However, it carries a higher risk of producing hallucinated or factually incorrect information. In our architecture, each model generates responses by conditioning on the user query, the retrieved posts, and the detected emotional tone. This pipeline ensures that the responses are not only contextually relevant but also emotionally aligned and grounded in peer support content.

Domain-Specific Questions:

To thoroughly test our Emotion-Aware RAG system, we created a variety of real-life mental health scenarios based on Reddit support posts. These cover a wide range of emotions like anxiety, sadness, fear, guilt, and optimism.

Each scenario acts as a test case to check if the system can correctly detect emotions, find similar past posts, and generate helpful and caring responses. This helps us see how well the system works across different emotional situations.

- **Anxiety (Sleep-Related)**
  Can the system support users who express restlessness and insomnia due to anxiety, offering calming and sleep-oriented responses?

- **Hopelessness (Suicidal Ideation)**
  Does the system safely handle emotionally critical posts that express despair or suicidal thoughts, prioritizing empathy and safety?
- **Fear (Career Instability)**
  Is the system capable of grounding responses with constructive reassurance when users express fear of losing their job?

- **Sadness (Breakups and Loss)**
  Can the model retrieve relatable content and generate supportive messages when users experience emotional grief after breakups?

- **Optimism (Recovery and Helping Others)**
  How well does the system acknowledge users who express recovery and willingness to help others, promoting resilience?

- **Anger (Unexplained Irritability)**
  Is the system equipped to handle intense emotional input and guide users safely through expressions of anger or frustration?

- **Loneliness (Social Isolation)**
  Can the system detect nuanced emotions like social disconnection despite physical presence and provide compassionate responses?

- **Guilt (Burden to Others)**
  Does the system offer reassurance and reduce self-critical thoughts when users feel like a burden to their friends or family?

- **Family Stress (Parent Conflict)**
  How effectively does the model address complex emotional inputs involving familial relationships and recurring anxiety?

- **Gratitude (Therapy Progress)**
  Can the system validate positive experiences, such as those shared by users who feel therapy is beginning to help them?

Implementation:

The Emotion-Aware Retrieval-Augmented Generation (RAG) system was implemented as a modular pipeline, integrating emotion classification, semantic retrieval, and response generation. We began with data collection using the Reddit Mental Health Support dataset sourced from Kaggle. This dataset contained thousands of user-generated posts discussing a range of mental health concerns. Preprocessing steps included converting text to lowercase, removing line breaks and special characters, and filtering out non-mental health-related content. The cleaned dataset was then used for both embedding and classification tasks.

For emotion detection, we employed the roberta-base-go_emotions model, which classifies input text into one of 28 fine-grained emotional labels such as sadness, fear, anger, guilt, and optimism. This emotional label plays a central role in guiding both the retrieval of relevant posts and the tone of the generated response.

To enable efficient retrieval of contextually similar posts, we generated dense vector embeddings using the all-MiniLM-L6-v2 model from the SentenceTransformers library. Each post was converted into a 384-dimensional vector, capturing its semantic and emotional content. These embeddings were indexed using FAISS (Facebook AI Similarity Search), allowing us to quickly retrieve the top-k most similar posts based on cosine similarity to the input query.

We then constructed prompt templates that combined the user's original query, the retrieved support posts, and the detected emotion. These structured prompts were fed into the response generation models. We used two large language models (LLMs) for comparison: **Phi-3 Mini**, known for its safety-aligned, empathetic outputs, and **Falcon-7B-Instruct**, which offered more fluent generation but showed higher risks of hallucination and repetition. Each model generated a personalized and emotionally sensitive response intended to provide comfort and actionable advice.

Finally, we evaluated system performance across multiple dimensions. Emotion detection was assessed using accuracy and F1 scores. Retrieval effectiveness was measured using Precision@k, indicating how well the retrieved posts aligned with the user's emotional state. For response generation, we used BLEU

and ROUGE (1, 2, and L) scores to quantify textual relevance and overlap. In addition, we conducted qualitative assessments focusing on empathy, safety, and hallucination frequency through manual review.

Comparative Results Across the LLMs:

Figure 2: Comparative Results Across the LLMs:

We compared the performance of two language models, Phi-3 Mini and Falcon-7B-Instruct, across several key metrics to evaluate how well they generated emotionally supportive responses. The evaluation focused on three main aspects: the quality of generated responses, their relevance to the retrieved content, and their emotional appropriateness for mental health support.

Phi-3 Mini demonstrated strong performance in generating safe and emotionally sensitive responses. It achieved a high Precision@k of approximately 83%, indicating that the top 3 Reddit posts retrieved for each query were highly relevant. For generation quality, Phi-3 scored moderately on BLEU (2.5) and

ROUGE-1 (0.18), which shows that its responses were reasonably similar to real-world reference replies but not overly focused on exact word matching. More importantly, the responses from Phi-3 were empathetic, context-aware, and free from hallucinations or unsafe suggestions. The model consistently offered supportive advice like encouraging users to talk to therapists or reach out to friends, making it ideal for use in mental health applications.

In contrast, Falcon-7B-Instruct achieved slightly higher BLEU scores (~3.1), suggesting better word overlap and fluency. However, this came at the cost of lower ROUGE scores (e.g.ROUGE-1 around 0.14), which indicates that while the language sounded more natural, the content was less grounded in the retrieved posts. Additionally, Falcon sometimes exhibited severe repetition and hallucination problems. In several cases, it repeated phrases like "I'm anxious and can't stop overthinking" more than 10 times in a single response. This made the replies feel robotic and potentially distressing to users. Although Falcon showed strong language capabilities, it lacked the emotional restraint and grounding needed for safe, personalized mental health assistance.

Overall, Phi-3 Mini outperformed Falcon-7B in emotional alignment, safety, and content grounding, while Falcon-7B showed stronger fluency and surface-level word matching. Given the sensitive domain of mental health, Phi-3 was clearly more reliable and suitable, proving that smaller, safer models can be more effective than larger models in emotionally sensitive AI applications.

Analysis:

Phi-3 Mini and Falcon-7B-Instruct showed clear differences across response quality, factual accuracy, reasoning, and empathy. Falcon-7B produced longer, more fluent, and conversational responses, often sounding natural but occasionally becoming overly verbose or repetitive—for example, repeating phrases like "I'm anxious and can't stop overthinking," which made the response feel robotic and less helpful. Phi-3 Mini, although not as fluent (lower BLEU score), generated shorter, focused, and emotionally appropriate replies that directly addressed the user's concerns without unnecessary repetition.

In terms of factual accuracy, Phi-3 Mini remained well-grounded in the retrieved Reddit posts, incorporating contextually relevant insights that built trust. In contrast, Falcon sometimes hallucinated information or introduced content not found in the retrieved context, which can be risky in sensitive mental health conversations. When evaluating reasoning and relevance, Phi-3 leveraged both the emotion label and retrieved content more effectively, offering empathetic and logically connected advice. Falcon's reasoning was broader but often less emotionally aligned with the user's specific needs.

Finally, in terms of empathy and safety, Phi-3 Mini consistently delivered supportive, actionable, and safe suggestions, such as recommending therapy or reaching out to trusted individuals, whereas Falcon's responses sometimes lacked emotional depth and, in a few cases, failed to respond safely to emotionally intense queries. Overall, Phi-3 Mini provided responses that were safer, more grounded, and better suited for emotionally sensitive applications like mental health support.

Strengths and Weaknesses of each model:

Phi-3 Mini

Phi-3 Mini demonstrated several key strengths that made it particularly suitable for emotionally sensitive applications like mental health support. One of its strongest qualities was its ability to generate emotionally aware and supportive responses. The model consistently produced replies that were context-sensitive, empathetic, and grounded in the user's detected emotion and the content of retrieved Reddit posts. This emotional intelligence made users feel understood, and the advice it provided was typically safe, encouraging, and actionable—such as suggesting reaching out to a therapist or trusted friend. Another major advantage of Phi-3 Mini was its efficiency and lightweight design. It required significantly less computing power and memory than larger models, allowing for faster response times and easier deployment on modest hardware like Google Colab or edge devices.

However, Phi-3 Mini also had a few limitations. While it was good at generating concise and safe replies, it sometimes lacked fluency and elaboration. Its responses could feel brief or overly simple, and it did not always provide the kind of detailed, nuanced reply a user might expect in a long conversation. It also scored slightly lower on BLEU metrics, which measure word overlap, because it focused more on emotional relevance than exact phrasing. Despite these limitations, Phi-3's strengths in emotional grounding, safety, and reliability made it a better fit for applications where user well-being is a priority.

Falcon-7B-Instruct

Falcon-7B-Instruct stood out for its linguistic fluency and natural-sounding responses. It produced longer, more detailed replies that often resembled human writing in tone and structure. This helped it achieve higher BLEU scores and made its outputs feel more polished in general conversations. Falcon's ability to follow instructions was also strong, allowing it to respond to a variety of prompts with flexibility and coherence. Its advanced language modeling capabilities made it suitable for general-purpose applications where fluency and verbosity are desirable traits.

Despite these strengths, Falcon-7B-Instruct also had notable weaknesses, particularly in the mental health support domain. It sometimes hallucinated content, inventing details not found in the retrieved Reddit posts. This created a risk of misleading or unsafe suggestions, especially in conversations dealing with emotional distress. Additionally, it suffered from repetition issues, with certain phrases being repeated multiple times in a single output, which made the conversation feel unnatural or robotic. Falcon's responses also tended to be less emotionally aligned, often missing the deeper emotional tone of the query. Moreover, it required significantly more computational resources, making it slower and harder to use in lightweight or real-time applications. These limitations made Falcon less dependable in emotionally sensitive contexts, despite its otherwise impressive language generation abilities.

Conclusion:

In this project, we developed an Emotion-Aware Retrieval-Augmented Generation (RAG) system designed to provide empathetic, safe, and contextually grounded responses for users seeking mental health support. The system combines emotion classification using RoBERTa-GoEmotions, semantic retrieval using MiniLM and FAISS, and response generation via two contrasting language models: Phi-3 Mini and Falcon-7B-Instruct. By integrating these components, we aimed to create a pipeline that could understand user emotions, retrieve emotionally aligned peer experiences from Reddit, and generate supportive responses tailored to the user's needs.

Our findings showed that Phi-3 Mini, though smaller in size, was highly effective in producing emotionally relevant and safe replies. It stayed grounded in the retrieved content and consistently provided actionable, caring suggestions. Falcon-7B, while more fluent, struggled with hallucinations and repetition, which made it less reliable in emotionally sensitive situations. These results reinforce the idea that emotional awareness, contextual grounding, and safety are often more important than raw fluency in applications like mental health support.

Overall, this project demonstrates the potential of combining emotion detection and retrieval-based generation with lightweight, responsible LLMs to build AI systems that are not only intelligent but also emotionally aligned and trustworthy. Such systems could play a meaningful role in digital mental health tools, especially when used alongside professional care.

References:

[1] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Sachan, D. (2020). *GoEmotions: A Dataset of Fine-Grained Emotions*. arXiv preprint arXiv:2005.00547. https://arxiv.org/abs/2005.00547

[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … & Riedel, S. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv preprint arXiv:2005.11401. https://arxiv.org/abs/2005.11401

[3] Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv preprint arXiv:1908.10084. https://arxiv.org/abs/1908.10084

[4] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. OpenAI Blog. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[5] OpenAI. (2023). *GPT-4 Technical Report*. OpenAI Technical Report. https://openai.com/research/gpt-4

[6] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … & Liu, P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. *Journal of Machine Learning Research*, 21(140), 1–67.

[7] Bhatia, A., Aggarwal, S., & Singh, M. (2023). *Hallucinations in Large Language Models: Causes, Impacts, and Mitigations*. arXiv preprint arXiv:2302.12153. https://arxiv.org/abs/2302.12153

[8] Wang, Y., Lin, K., & Wang, B. (2022). *Safety Challenges in Conversational AI: A Survey*. *ACM Computing Surveys (CSUR)*, 55(9), 1–38.

[9] Touvron, H., Lavril, T., Izacard, G., et al. (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv preprint arXiv:2302.13971.

[10] Microsoft Research. (2024). *Phi-3 Technical Card*. Hugging Face Model Card. https://huggingface.co/microsoft/phi-3-mini-128