ECE 7650-T39: Deep Generative Modelling **Winter 2025**

Project Report
Adaptive Retrieval-Augmented Generation for Psychological Assessment

Muhammad Adeel Zahid *April 04, 2025*

Learning Outcomes

- Understand the core principles behind retrieval-augmented generation (RAG) and its application in psychological assessment.

- Learne how to use text embedding models such as Sentence Transformers and Hugging Face Transformers for semantic similarity tasks.

- Gain insight into adaptive k*-nearest neighbor algorithms for dynamically selecting relevant user-generated content.

- Gain experience with the process of extracting mental health insights using clinical questionnaires like GAD-7 and PHQ-9.

- Learn how to structure effective prompts for large language models (LLMs) to elicit diagnostic responses.

- Acquire practical skills in building and debugging LLM-based pipelines in Python using PyTorch, tensorflow and NumPy.

Problem Statement

Mental health disorders like depression are often underdiagnosed due to stigma, limited access to care, and the inherent difficulty of accurately interpreting and processing complex human emotions. Understanding a person's psychological state is a nuanced task, making it challenging to reliably assess conditions like depression through conventional means. Reliance on formal self-report tools such as GAD-7 **(**Generalized Anxiety Disorder 7-item**)** may not capture real-life emotional experiences. Meanwhile, platforms like Reddit contain rich, unstructured expressions of mental health struggles. This project proposes a Retrieval-Augmented Generation (RAG) pipeline that uses Large Language Models (LLMs) and adaptive embedding-based retrieval to infer psychological assessment scores such as GAD-7 from Reddit posts. Text fetched by this pipeline is presented to the LLM with a prompt specifying an item from GAD-7 such as "Feeling nervous, anxious, or on edge". LLM then outputs a score along with relevant explanation. By applying zero-shot prompting and structured templates, the system maps user-generated content to standardized mental health scales, allowing scalable and context-aware mental health inference.

Background Theory

Traditional psychological assessments like the GAD-7 [1] rely on structured self-report questionnaires, which, while clinically validated, require respondents to explicitly reflect on their mental state, something many individuals may avoid or struggle at it. On the other hand, social media platforms such as Reddit offer naturally occurring expressions of thoughts, emotions, and behaviors, providing an alternative lens to users' mental health.

In recent years, social media platforms like Reddit have emerged as spaces where users openly share their thoughts, emotions, and behaviors. The subreddit r/mentalhealth, for example, serves as a central forum for discussions related to mental well-being [2]. This unstructured user-generated content offers a rich data source for mental health research, providing insights into individuals' experiences outside traditional clinical environments. Studies have utilized Reddit data to explore various aspects of depression and anxiety, highlighting the platform's potential for mental health surveillance and intervention [3].

Large Language Models (LLMs) such as GPT-4 and Mistral-7B have demonstrated strong capabilities in understanding and generating human-like text, making them suitable for interpreting unstructured mental health discourse. However, fine-tuning such large models is often too expensive. Training GPT-4, for example, has been estimated to cost over $100 million [4]. Even fine-tuning smaller models like LLaMA-13B can require substantial computational resources, such as multiple high-end GPUs, which are not accessible to most researchers [5]. These constraints make zero-shot or prompt-based approaches more practical and scalable for many applications.

To effectively utilize LLMs in zero-shot settings, relevant content must first be selected from a user's extensive Reddit history. Semantic embeddings provide a way to measure the similarity between assessment items and Reddit posts. Instead of using a fixed number of neighbors, adaptive retrieval methods like the k*-nearest neighbor algorithm estimate the optimal neighborhood size based on the data's local intrinsic dimensionality [6][7][9].

Recent work has explored combining language models and retrieval methods for psychological assessment using social media data. Lewis et al. [8] proposed a framework, which improves factuality by grounding LLM responses in retrieved documents. Su et al. [9] demonstrated that large language models like GPT-4 can approximate clinician-rated psychological assessments in zero-shot settings. Valdez et al. [10] applied zero-shot classification to social media posts for mental health symptom detection, highlighting the scalability of prompt-based inference.

Methodology

To address the challenge of mapping informal Reddit posts to standardized psychological assessments, a Retrieval-Augmented Generation (RAG) pipeline is implemented, composed of three main components: document embedding, adaptive retrieval, and large language model (LLM)-based inference.

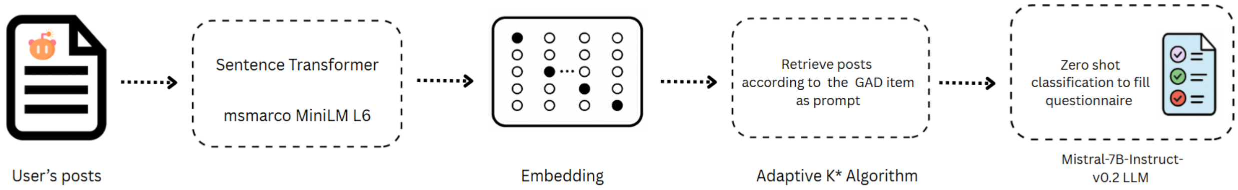
Sentence Transformer
msmarco MiniLM L6
Retrieve posts according to the GAD item as prompt
Zero shot classification to fill questionnaire
User's posts
Embedding
Adaptive K* Algorithm
Mistral-7B-Instruct-v0.2 LLM

Figure 1: RAG pipeline

User's reddit posts are downloaded from reddit using reddit's API and are subsequently converted into embeddings using sentence transformer called msmarco-MiniLM-L6-cos-v5 [11]. Different reddit posts may be downloaded depending upon the time of day. A reference json file is provided to replicate the same results as ours. The main LLM is loaded called Model Card for Mistral-7B-Instruct-v0.2 [12].

Adaptive Semantic Retrieval via K* Neighborhoods

To identify the most contextually relevant Reddit posts for each psychological assessment item, we utilize an adaptive nearest-neighbor retrieval mechanism based on the intrinsic dimensionality of the data. The implementation of the retrieval algorithm in this project has been adapted based on prior research by Facco et al. [6] and Su et al [9]. This method is implemented in the code through the functions *compute_kstar_with_adaptive_method* and *find_single_k_neighs*.

Embedding Construction

Each questionnaire item is first embedded into a fixed-dimensional vector space using the SentenceTransformers model. Similarly, all Reddit posts by a user are embedded into the same vector space, forming a matrix X ∈ Rn × d, where n is the number of posts and d is the embedding dimension.

K* Estimation

We concatenate the item embedding with the user's post embeddings and construct a new dataset X' = [xo;X] ∈ R(n+1)×d, where xo is the questionnaire item vector. The intrinsic dimensionality D of this dataset is estimated using the two-nearest-neighbors (2NN) estimator, followed by the estimation of the optimal neighborhood size k∗ using a binomial model

$$k^* = \arg\max_k \mathcal{L}(D \mid k, n, r)$$

Where '*L*' is the binomial log-likelihood defined as:

$$\mathcal{L}(D) = \sum_{i=1}^{n} \left[n_i \log(p) + (k_i - n_i) \log(1-p) \right], \quad p = r^D$$

The search for k∗ is repeated for a fixed number of iterations to stabilize the dimensionality estimate. This process can be found in the function *compute_kstar_with_adaptive_method*.

Cosine Similarity-Based Retrieval

After determining k*, the k* most relevant Reddit posts to the item embedding are retrieved by calculating cosine distances between the item and all post embeddings

$$\text{cosine\_distance}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

The indices of the k∗ posts with the smallest cosine distances are returned by the function *find_single_k_neighs*, and these posts are passed to the LLM for scoring the user's psychological state on that item.

Prompt Construction

The retrieved posts are truncated to fit the LLM context window and passed into a prompt template for inference. The LLM is also instructed to report the score for each item. In the provided JSON file, there are reddit posts for four users. There are five items for which a score is calculated such as "Feeling nervous, anxious, or on edge", "Not being able to stop or control worrying" and so on. For each of these items relevant reddit posts or pieces of text are pulled and combined with in a prompt given to the LLM. LLM then gives a score for each user for that item on a scale of zero to five. Zero means that the occurrence of this item is "not at all" and goes on to "several days", "more than half of days" or "nearly every day" which corresponds to score of five. The prompt template is carefully constructed to guide the large language model (LLM) through a chain-of-thought (CoT) reasoning process. By explicitly instructing the model to reflect step-by-step, first analyzing user-generated Reddit content, then selecting the most appropriate response option, and finally justifying and scoring it. The model is encouraged to engage in interpretable and structured reasoning. This not only improves the quality of the predictions but also helps mitigate hallucinations by grounding the LLM's output in retrieved evidence. The final instruction to output the score in a strict format (SCORE_OUTPUT) ensures response consistency and facilitates automated extraction of the final diagnostic signal. The fixed prompt is as given below. There are tags for reddit posts which is the retrieved and filtered text, item names and choices which are inserted dynamically to form the complete prompt. Here choices refer to the zero-to-five scale in words as described above.

Fixed Prompt:
"You are a psychological assistant specializing in administering and interpreting standardized psychological assessments.
Your task is to assist in filling out the Generalized Anxiety Disorder 7-item scale (GAD-7), a widely used self-report questionnaire designed to assess the severity of generalized anxiety disorder, based on the user's Reddit posts.

Step A: Consider the following Reddit posts: {REDDIT_POSTS}.
Step B: Based on the relevant Reddit posts for answering a question related to {ITEM_NAME}, choose which of the following choices seems most appropriate as a response: {CHOICES}.
Why does this choice stand out as the best match given the user's current psychological state? Think about the reasoning behind this choice step by step.
Step C: Finally, report the final score (0-3) based on the intensity of {CHOICES}. Use the reasoning from the previous steps to evaluate your scoring. Write in the fixed format SCORE_OUTPUT:?"

Results

For brevity, one psychometric item is discussed from the questionnaire across two selected users. The rest of the responses can be referred from the attached "*responses.txt*".

Item: Feeling nervous, anxious, or on edge

User 01:
Relevant Reddit Posts/text:

- Can someone help me answer this question? So I don't think I'm bipolar but I can go from not being suicidal to being suicidal and back to not suicidal very fast Like it can happen rly fast like when I'm not suicidal it feels like I was never suicidal and I never will be but when I am it feels like I always was and always will be I think I need to go to a mental hospital but does any1 have any advice¿
- This is the first time I've ever been suicidal but didn't want to talk to someone or even feel better at all, everyorher time it happened I wanted to talk it out with someone or do something I like but this time it's different
- Ye I've been on the brink of making a plan for the past few months but I've always just about managed to distract myself in any way I could so I didn't start making a plan, the closest I got was I started writing notes for people I know incase I did it
- No I can't, I have an intense phobia of hypnosis from the last time I did it cuz of what it made me do.

LLM Response:
"Step A: The user's posts suggest a history of suicidal ideation and mood instability, which are symptoms of both anxiety disorders and bipolar disorder. However, the user expresses uncertainty about their diagnosis and does not explicitly mention symptoms of mania or hypomania, which are key features of bipolar disorder. Therefore, based on the information provided, it seems more likely that the user is experiencing symptoms of anxiety disorder, specifically generalized anxiety disorder (GAD).

Step B: The user's posts indicate that they have experienced periods of intense anxiety and suicidal ideation that come and go rapidly, and that during these periods they feel consumed by their anxiety and despair. This suggests that they experience anxiety nearly every day, as described in choice 3.

Step C: Based on the user's self-reported symptoms, a score of 3 on the GAD-7 scale would be most appropriate. This score indicates that the user experiences anxiety nearly every day, which aligns with the user's description of their psychological state.

SCORE_OUTPUT: 3."

User 02:

Relevant Reddit Posts/text:

- Yeah but I feel like it's excessively bad, gets to the point I can't even be around people for a moderate amount of time.
- Is it bad that I have no desire to make any friends? I literally don't want to put in the effort to make friends. Is that bad of me? I feel fine most of the time like this.
- Is it weird that I've been alone so much I actually enjoy it more than hanging out with people. I've gotten to the point I actually enjoy doing stuff all alone and would choose it over hanging out with friends and stuff. Don't know what has happened but I feel more at peace.
- Not even interested in my degree anymore No part of me is interested in my degree anymore. I've lost complete interest in studying and anything to do with this. I'm just trying to make it through to summer but everytime I sit down and try to do work I end up just frustrated and angry. Fuck this life.
- Yeah, and its even with close friends too.
- I have friends, just don't want to make new ones really.
- Is life genuinely worth it if 80% of the time is suffering and 20% is actual rewarding fun that feels fulfilling? I never understood why we live a life where we suffer considerably more than we actually have fun. I read about people online and see people in real having to suffer, make sacrifices just to be able to scrape by and live on. I understand I am more privileged than lots of people in this world, but even then it doesn't seem worth it if I just slave away and suffer. I've already given up on so many things I was once interested in because I just don't have the time or energy to do them anymore. What's the point if surviving is just suffering?.

LLM Response:

"Step A: The user's Reddit posts indicate a pervasive sense of anxiety, avoidance of social situations, lack of interest in activities, and a general sense of dissatisfaction with life. These symptoms are consistent with those observed in individuals with Generalized Anxiety Disorder (GAD).

Step B: Based on the user's descriptions of feeling unable to be around people for extended periods, having no desire to make friends, and preferring to be alone, it appears that the user experiences anxiety or nervousness more than half the day. This is consistent with a GAD-7 score of 2.

Step C: SCORE_OUTPUT: 2. The user's symptoms suggest a moderate level of anxiety and nervousness, which corresponds to a GAD-7 score of 2."

The presented outputs illustrate how the LLM processes relevant Reddit posts to estimate anxiety severity using the GAD-7 framework. For User 01, the model identified consistent, intense expressions of suicidal ideation and emotional distress, leading it to assign the maximum score of 3, indicating symptoms occurring nearly every day. In contrast, User 02 displayed signs of social withdrawal, emotional fatigue, and existential dissatisfaction, which the model interpreted as moderate anxiety symptoms, resulting in a score of 2. These outputs highlight the model's ability to reason over user-generated text and assign clinically aligned severity levels based on the user's language patterns and affective expressions.

Summary

This project demonstrates an approach to automated psychological assessment using large language models (LLMs) and adaptively retrieved Reddit posts. Relevant user-generated content is first embedded using SentenceTransformers, and item-specific retrieval is performed using the K* neighborhood algorithm, which selects posts based on intrinsic dimensionality and proximity. Prompts are constructed to guide the LLM (Mistral-7B-Instruct-v0.2) through step-by-step reasoning for each assessment item. The output is a predicted score that aligns with standardized scales such as the GAD-7. This method is designed to encourage interpretability, reduce hallucinations, and ensure that the model's outputs are grounded in the user's own text.

References

1. Löwe, B., et al. (2008). "Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population." *Medical Care*, 46(3), 266-274. LinkPubMed
2. Reddit. r/mentalhealth. LinkReddit
3. Gkotsis, G., et al. (2017). "Characterisation of mental health conditions in social media using Informed Deep Learning." *Scientific Reports*, 7, 45141. LinkPMC
4. Wikipedia contributors. "GPT-4." *Wikipedia, The Free Encyclopedia*. LinkWikipedia
5. OpenAI Developer Community. "GPT-4 Cost Estimate (UPDATED)." LinkOpenAI Community
6. Facco, E., d'Errico, M., Rodriguez, A., & Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1), 12140.
7. Sricharan, K., & Hero, A. O. (2010). "Optimized intrinsic dimension estimation using nearest neighbor graphs." *IEEE International Conference on Acoustics, Speech and Signal Processing*. Link
8. Lewis, P. et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS.
9. Su, Y. et al. (2024). Are LLMs Effective Psychological Assessors? arXiv:2501.00982.
10. Valdez, D. et al. (2021). Zero-Shot Text Classification for Mental Health Surveillance. JIS.
11. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Hugging Face*, [Online]. Available: https://huggingface.co/sentence-transformers/msmarco-MiniLM-L6-cos-v5.
12. Mistral AI, "Mistral-7B-Instruct-v0.2", *Hugging Face*, [Online]. Available: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2.