Project Group 04: Discovering Topics in Dialectic Behavioral Therapy through LLM

Authors

Jesse Parent: jeparent@ucsd.edu

Nathaniel Do: ntd002@ucsd.edu

Abstract Details

Background

We are working on identifying key words and phrases associated with dialectic behavioral therapy (DBT), a psychotherapy focused on managing emotions, improving relationships, and reducing self-destructive behavior. Our analysis will incorporate large language models (LLMs) to analyze text from Reddit's subreddit (a social media community) r/dbtselfhelp. We have been in contact with Dr Wiebel and his team from the HXI lab at UCSD to discuss their forthcoming project within this problem space.

Problem Definition

Can we build a classification model to analyze user posts from r/dbtselfhelp and predict the underlying topics, such as anger, guilt, or acceptance? The model will take raw text as input and output a categorized topic or keyword, enabling further exploration of mental health-related content. (Note: this is a potential framing of an initial problem set for a general scope, the larger problem of an AI agent fluent in DBT would have other, more complex goals)

Motivation & Lit Survey

In terms of broader motivation, mental health is a difficult topic in terms of both access, stigma for treatments, and difficulty matching available, trained, and/or licensed clinical social workers and psychologists to the needs and circumstances of various individuals. One way of addressing this challenge is through creating more accessible AI-based agents who can play a preliminary or auxiliary role in an individuals treatment. LLMs are popular and face a rush of contemporary applications, particularly in terms of extending the reach of existing health and mental health services.

We seek to use ML to identify keywords associated with various mental health areas and to find patterns in the posts. The subreddit has over 41 thousand members with a few posts per day since 2012, leading to a large text dataset for us to use. This is also an ongoing task by Dr. Weibel's team of DSC 266. Their desire is to build a retrieval mechanism to provide responses to posts of detected topics. We will also incorporate techniques learned from our time at UCSD like word pairing, stemming, and discarding. By combining these approaches and adjusting the values, we aim to find different results than they receive.

Dataset

Dr. Weibel of the HXI lab at UCSD has supplied us with this initial cache of material, which contains posts from the reddit thread, documents and several associated texts that offer help regarding mental health support, and other materials. The given data will need to be evaluated in terms of its breadth and balance of topics. Some of the "ground truth" may be which posts were posted under which category or subforum within the reddit main forum. There will need to be a determination of whether or not the given data will provide enough means for solid training, or if we will need to go elsewhere to develop a model before training on that data itself.

Approach

We will explore a variety of options in order to discern what is most functional for this kind of task. We will start with a general sentiment analysis, and may also include:

- Sentiment analysis using pre-trained models like BERT or GPT.
- Topic modeling via Latent Dirichlet Allocation (LDA) or non-negative matrix factorization (NMF).
- Classification using SVM, logistic regression, or fine-tuning transformers.

As per Dr Weibel's team, their end goal may involve using Retrieval-Augmented Generation (RAG) as a means to generate a "bot" or some form of agent /agentic AI to interface with those seeking mental health support. In our time for this Capstone, we may explore the usefulness of RAGs, but a more preliminary step may be sentiment analysis and classification of the content of user's posts on the Reddit forum.

Statement on Existing Solutions: there is no specific set solution for this problem, so our work is exploratory in nature. The HXI lab is in the process of exploring this problem space and these resources themselves, and our efforts are complementary to their process.

Success Criteria

A varying scale of success, given our broader scope. A first step may be simple classification and matching a post's content to the title or subforum it was in. But more useful outcomes may involve generating a topic, problem, or lead that otherwise an AI agent could center on and direct further conversation around; in this way, our work may be seen as preliminary scouting or generating a baseline for later developments in this project.

- Be able to identify a post's topic. For example, does our model perform better than random chance, or a very simple SVM or Bag of Words model?
- Evaluate and compare what models work best for generating meaningful outcomes
- A bonus goal may be to implement a RAG itself and compare or build on top of our initial work to further assist in the Weibel Lab's efforts: implementing a preliminary RAG might work well for topic-based responses as they can pull contextually relevant information from external datasets, which were supplied in the original cache.

Challenges:

There are also challenges we will be navigating, including:

- User bias in how content is generated in an online forum

- Privacy: privacy access to Reddit and scraping data from it is challenging and has its own obstacles. We have had a preliminary discussions with Dr Weibel’s team about this, and may have more over the coming weeks
- Research and exploratory investigations: we’re taking a bit of a risk here by trying to venture into relatively unknown territory, which may shape some of our outcomes compared to other projects with established datasets and problem spaces. But we are ok with this as an endeavor, and have attempted to communicate about this with the Course Instructor and also Dr Weibel and his team.