Leveraging LLMs to Understand Global Mental Health Well-being & Fomo

INTERACTIVE HELP TO ALLEVIATE MENTAL HEALTH ISSUES DUE TO SOCIAL MEDIA USAGE FOCUS ON SOCIAL MEDIA ADDICTION, SOCIAL ISOLATION AND CYBERBULLYING

Challenge Hosted by: OMDENA HYDERABAD CHAPTER, INDIA

March 26, 2024

TABLE OF CONTENT

INTRODUCTION

In the realm of mental health, grasping the interconnected nature of well-being is paramount, especially within the evolving landscape of the knowledge economy. The recognition of cognitive and emotional skills as crucial extends beyond professional spheres to empower a diverse and remote workforce.

However, amid these advancements, the prevalence of social media usage introduces challenges such as social media addiction, social isolation, and cyberbullying, which significantly impact mental health. Our project is dedicated to addressing these challenges directly by offering personalized strategies aimed at providing actionable guidance to individuals on their unique paths toward mental well-being.

This pioneering initiative aligns seamlessly with contemporary society's nuanced understanding of mental health, marking the advent of a transformative era in mental health care.

Summary

As social media deeply impacts human psychology, our goal was understanding both positive connectivity and negative consequences like social media addiction, isolation and cyberbullying. With Al growing rapidly, we developed an innovative LLM chatbot using Direct Preference Optimization.

This chatbot provides personalized mental health support, fostering self- reflection and community engagement. Our human-centric approach addresses the nuanced challenges modern humans face in navigating social media's complexities while prioritizing holistic well-being.

Using a fine-tuned Mistral 7b LLM model, trained on therapy-like counseling responses, we have developed a chatbot that knows how to properly respond to various mental health issues that a user may experience.The chatbot is given extensive knowledge in order to provide the user with the resources and answers to any question they may have on mental health issues related to social media.

Data for training consisted of user queries from social media and online counseling conversations, as well correct and incorrect responses to those queries generated from a general LLM. Then with DPO fine-tuning our model is able to give more appropriate and relevant responses. Sentiment analysis was also conducted on the data and is integrated into our project to evaluate the mood of the user and respond accordingly.

The UI will feature a questionnaire apart from the chatbot to find areas to focus on in conversation with the user. Voice-to-text can also be used to allow for more fluent conversation with the chatbot. These features allow for a more personalized experience for the user and will have their privacy in mind with top security features to keep all conversations and information private and only seen by the user. Finally, the chatbot is deployed on the Google Cloud Platform and is available to provide help to people globally.

Task

Omdena Hyderabad Local Chapter brought together 33 collaborators from countries around the world. Collaborators were tasked with conducting research on mental health issues associated with social media addiction, social isolation, and cyberbullying stemming from social media usage.

The team was required to devise a design approach that would shape the criteria for data collection and model tuning. The chosen design approach was Directed Preference Optimization (DPO), necessitating the use of preference pairs for data collection.

The teams responsible for data collection, data preprocessing, model development, and UI design collaborated iteratively and synergistically to create an interactive, cutting-edge solution. Leveraging a DPO-trained model, the solution offers insightful guidance to users, aiding them in addressing the mental health challenges arising from their social media usage.

Problem Statement

In today's digital landscape, social media addiction, social isolation, and cyberbullying have emerged as pressing concerns, warranting focused attention. With platforms like Facebook, Twitter, TikTok, SnapChat, WhatsApp and Instagram deeply embedded in daily life, the detrimental effects on mental health have become increasingly apparent.

Social media addiction grips individuals in a relentless cycle of compulsive usage, often leading to neglect of real-world relationships and responsibilities. This dependency can exacerbate feelings of isolation as users retreat further into the digital realm, diminishing face-to-face interactions and genuine human connections.

Simultaneously, social isolation can be both a cause and consequence of excessive social media use. The allure of virtual communities may initially offer a sense of belonging, but over time, it can foster a sense of disconnection from the offline world, amplifying feelings of loneliness and alienation.

Furthermore, the anonymity and ubiquity of social media platforms provide fertile ground for cyberbullying, a pervasive and damaging phenomenon. The anonymity offered by online interactions emboldens perpetrators to engage in harmful behaviors, causing profound psychological distress and trauma for victims.

Addressing these interconnected issues requires a multifaceted approach that encompasses both individual and systemic interventions. Educating users about healthy digital habits, promoting digital literacy, and fostering supportive offline communities are essential steps in mitigating the negative impact of social media on mental health.

By acknowledging the complex interplay between social media addiction, social isolation, and cyberbullying, society can work towards creating a healthier digital environment where individuals can thrive both online and offline.

Goals and Objective

The project goals include:

- To conduct an in-depth analysis of the impact of social media on various aspects of mental health, including social media addiction, social isolation and cyberbullying
- Develop targeted interventions and recommendations to mitigate negative mental health effects associated with social media use.
- Provide insights into fostering a healthier online environment, promoting positive interactions, and leveraging social media for mental well-being.

Market Research Analysis

- **Industry Overview:**
  The $435.2 billion mental health market is growing, driven by increasing demand, digital innovation, corporate wellness initiatives, and supportive policies. This creates opportunities for leveraging large language models (LLMs) and retrieval-augmented generation (RAG) approaches.

- **Target Market:**
  Teenagers (16-20) struggling with mental health issues exacerbated by social isolation, social media addiction, and cyberbullying.

- **Opportunities:**
  - Integrating LLMs and RAG to develop virtual mental health assistants offering scalable, personalized support after fine-tuning on relevant data.
  - Leveraging LLM embeddings and machine learning for predictive analytics and social media analysis to detect mental health risks early.

- **Working Landscape:**
  - Utilizing LLMs for automatic evaluation of virtual mental health counseling.
  - LLM/RAG systems should augment human experts, not replace therapists' critical diagnostic, treatment planning, and empathetic support roles.
  - Ongoing research aims to enhance LLM performance using online text data, social media analysis, and automated counseling evaluation.

- **Ethical Considerations:**
  - Ensure transparency, fairness, anti-bias measures, and compliance with data privacy regulations like HIPAA/GDPR.
  - Implement stringent testing, human oversight, informed consent processes, and robust safeguards.
  - Mitigate risks of inappropriate or misleading Al-generated mental health advice.
  - Collaborate with mental health professionals and follow ethical Al guidelines.

These AI systems could increase access to personalized mental healthcare, enable early intervention through social media analysis, and improve virtual counseling quality while complementing human expertise.

DATA COLLECTION

The data collection phase stands as a pivotal stage in our project, systematically gathering pertinent information to fulfill specific objectives. This phase entails the meticulous and organized acquisition of data from diverse sources.

Through extensive collaboration and thorough research, our team identified the essential key data elements to be captured during data collection and common threads/themes across different job positions. This process allowed us to refine our interview approach, ensuring that a generic design could be put in place to capture the necessary data elements effectively.

In our data collection endeavors, we capitalized on the capabilities of Large Language Models (LLMs) for generating extensive datasets. LLMs, such as GPT (Generative Pre-trained Transformer), have been pre-trained on vast amounts of text data, enabling them to understand and mimic human language patterns effectively. Leveraging LLMs for data generation offers scalability and versatility, allowing us to augment existing datasets, simulate scenarios, and produce domain-specific data efficiently. Incorporating LLM-generated data enriches our dataset with diverse examples, enhancing the robustness and performance of our machine learning models

For our project focusing on mental health, cyberbullying, and social isolation, we employed a multi-faceted approach to gather relevant data. We utilized advanced techniques such as prompt engineering to generate data from Gemini and Chatgpt, in JSON format. The collected data follows the format:

{

"prompt": "This is the user-given description of their situation which is evaluated.",

"chat_response": "This contains a chat response which the model deems as a correct evaluation.",

"rejected_response": "This contains a rejected response which the model deems as an incorrect evaluation."

}

This standardized format ensures consistency in the data collected, facilitating streamlined processing and analysis.

In addition to generated data, we leveraged the Reddit Mental Health Dataset, which comprises posts from various users struggling with mental health issues. This dataset provided valuable insights into real-world experiences and sentiments shared within online communities, enriching our understanding of the challenges individuals face regarding mental health, cyberbullying, and social isolation.By combining data from Gemini, ChatGPT, and the Reddit Mental Health Dataset, we ensured a comprehensive and diverse pool of information for training our chatbot model. This approach enables our chatbot to provide empathetic and informed responses to users seeking support and guidance on mental health-related issues.[2]

DATA PRE-PROCESSING

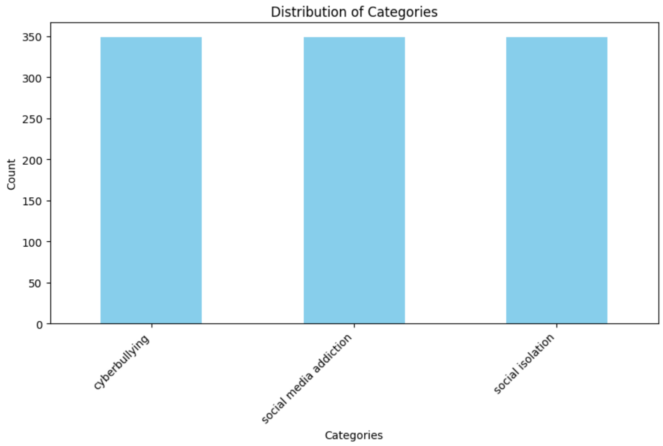
Data preprocessing is a crucial step in preparing a dataset for analysis and modeling tasks. This report outlines the comprehensive preprocessing procedures applied to a textual dataset focusing on expressions of social isolation. The dataset comprises prompts expressing feelings or situations, along with corresponding chosen and rejected responses, categorized under themes such as social isolation, social media addiction, and cyberbullying. The preprocessing methods described herein aim to standardize, clean, and enrich the text data to facilitate meaningful insights into the chosen themes.

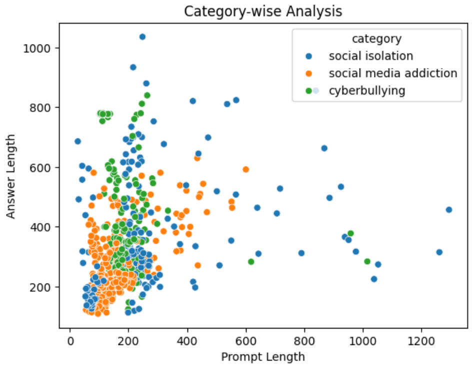
Data Preprocessing Methods:

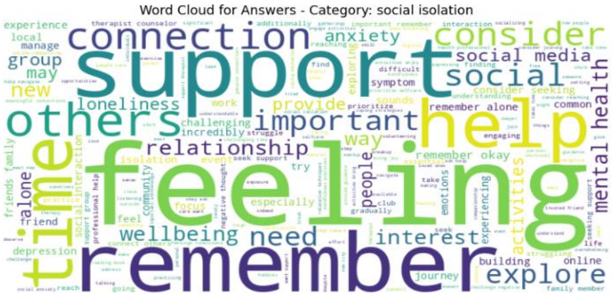
- **Text Cleaning:** Special characters, punctuation, and whitespace were removed, and text was converted to lowercase for consistency.
- **Handling Missing Values:** Missing values were examined, and decisions were made regarding imputation or removal based on their impact on the analysis.
- **Text Tokenization:** Prompts and responses were tokenized into individual words or tokens.
- **Stopword Removal:** Common stop words were removed to eliminate noise in the text data.
- **Feature Engineering:** Additional features such as word count and sentiment scores were extracted to enrich the dataset.
- **Standardizing Categories:** Different variations of category names were standardized to maintain consistency. For instance, 'SOCIAL ISOLATION', 'Social Isolation', and 'Social isolation' were replaced with 'social isolation'. Similarly, variations of 'SOCIAL MEDIA ADDICTION' and 'CYBERBULLYING' were standardized.
- **Removing Rows:** Rows with irrelevant categories were removed from the dataset to focus on the predefined themes related to social isolation, social media addiction and cyberbullying without overlapping into other mental health categories.
- **Removing Double Quotes:** Double quotes were removed from all columns in the DataFrame to ensure consistency and cleanliness of the text data.
- **Removing Duplicate Rows:** Duplicate rows in the DataFrame were identified and removed based on the combination of values in the 'prompt', 'chosen', and 'rejected' columns. This step ensures data integrity and eliminates redundancy in the dataset.
- **Text Length Calculation:** The lengths of the text in the 'prompt', 'chosen', and 'rejected' columns were calculated and stored in separate columns ('prompt_len', 'chosen_len', and 'rejected_len', respectively).
- **Category-wise Analysis:** A scatter plot was generated to visualize the relationship between the length of prompts and rejected answers, with each point colored by category for category-wise analysis.
- **Word Cloud Generation:** Word clouds were generated for each category to visualize the most frequent words in prompts, chosen responses, and rejected responses.
- **Top Words Analysis:** Bar plots were generated to visualize the top 10 most frequent words in prompts, chosen responses, and rejected responses for each category.

The preprocessing steps outlined in this report collectively contribute to the integrity, cleanliness, and suitability of the dataset for subsequent analysis and modeling tasks. By systematically addressing various data cleaning, standardization, and enrichment tasks, the dataset is optimized for accurate and meaningful insights into expressions of the task.

Additionally, EDA was an iterative process and heavily engaged and interactive with the data collection. Iterative analysis resulted in the culmination and subsequent confirmation of a balanced data set for use in our DPO model training:







Data Storage

After preprocessing, the dataset has been saved in CSV format for ease of access and compatibility. The CSV file, along with any associated files such as metadata documentation and analysis reports, is securely stored in two primary locations: Google Drive and DagsHub.

- Google Drive: The preprocessed dataset CSV file is stored securely on Google Drive. This platform provides convenient access to team members and ensures data integrity and availability.
- DagsHub: DagsHub serves as a collaborative platform for version-controlled data science projects. We have uploaded the preprocessed dataset CSV file to DagsHub, facilitating version tracking, reproducibility, and collaboration among team members.

Additionally, the Jupyter notebooks containing exploratory data analysis (EDA) insights, visualizations, and analysis scripts are stored alongside the dataset CSV file on DagsHub. This ensures that all analyses and visualizations are captured within the project environment, allowing for comprehensive documentation and reproducibility of results.

Using both Google Drive and DagsHub, we ensure that the preprocessed dataset and associated artifacts are securely stored, version-controlled, and readily accessible to authorized stakeholders for further analysis and research purposes.

CHATBOT UI DESIGN

The Social Media Addiction Assessment and Guidance App offers a user-friendly solution to address this concern, providing personalized guidance through an intuitive interface.

Upon launching the app, users navigate through a visually appealing questionnaire, answering prompts about their social media habits and experiences. Leveraging Python and the Streamlit library, the app collects and processes these responses, employing a pre-trained model to generate a tailored assessment.

The app's unique strength lies in its ability to provide personalized recommendations based on the assessment results and an additional user query. Users receive valuable insights and strategies to cultivate a healthier relationship with social media, promoting mental well-being.

Integrating a suitable pre-trained model, ensuring an intuitive user experience, and implementing robust data privacy measures were key challenges faced during development. With its user-friendly design, personalized assessments, and focus on mental health, the Social Media Addiction Assessment and Guidance App stands as a valuable resource for individuals seeking to understand and overcome the potential negative impacts of excessive
social media usage.

Question Generation/User Interaction

As a user embarks on this transformative journey, they are warmly welcomed into the app's intuitive and visually appealing interface. Upon launching, the user is prompted to provide their name and age, ensuring a personalized experience tailored to their unique circumstances.

The user then navigates through a meticulously crafted questionnaire, thoughtfully designed to explore their social media habits, experiences, and potential areas of concern. The user-friendly format, coupled with clear prompts and predefined answer options, fosters an engaging and comfortable experience, encouraging honest and thoughtful responses.

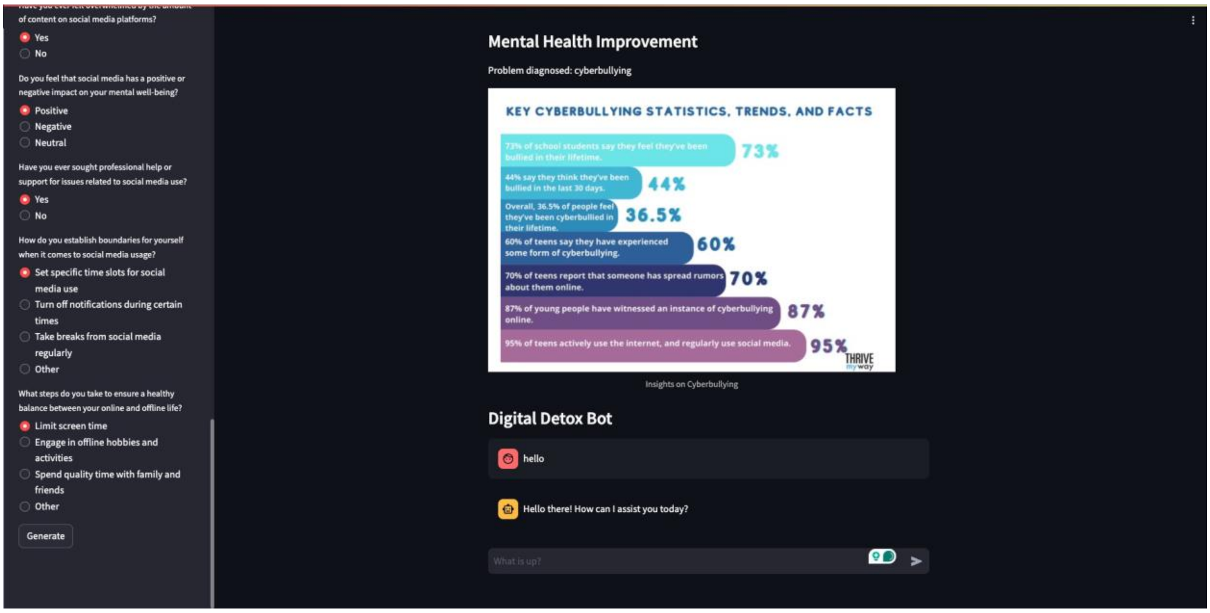
Once the form is submitted, the app seamlessly initiates the process of loading a sophisticated pre-trained mental health model. Leveraging advanced machine learning techniques, this model is capable of analyzing complex patterns in the user's responses contributing to the generation of accurate and insightful assessments.

The user is then presented with an opportunity to provide an additional query, allowing the app to further personalize the assessment process and address their specific concerns or areas of interest.

Combining the user's responses from the questionnaire, the additional query, and the loaded mental health model, the app generates a tailored assessment using the social addiction assessment and generates text functions. This assessment is then prominently displayed, offering the user valuable insights, recommendations, and actionable strategies to improve their mental well-being related to social media usage.

Throughout this interactive journey, the app maintains a visually appealing interface and a user-centric approach, ensuring a comfortable and empowering experience while leveraging cutting-edge technology to provide personalized guidance tailored to the user's unique circumstances. It is important to note that the app's boundaries are defined by the scope of the pre-trained model's capabilities and the depth of the assessment questionnaire. One potential drawback is the reliance on the accuracy and completeness of the user's responses.

Digital Detox UI Screenshot



Have you ever felt overwhelmed by the amount of content on social media platforms?

● Yes
○ No

Do you feel that social media has a positive or negative impact on your mental well-being?

● Positive
○ Negative
○ Neutral

Have you ever sought professional help or support for issues related to social media use?

● Yes
○ No

How do you establish boundaries for yourself when it comes to social media usage?

● Set specific time slots for social media use
○ Turn off notifications during certain times
○ Take breaks from social media regularly
○ Other

What steps do you take to ensure a healthy balance between your online and offline life?

● Limit screen time
○ Engage in offline hobbies and activities
○ Spend quality time with family and friends
○ Other

Generate

Mental Health Improvement

Problem diagnosed: cyberbullying

KEY CYBERBULLYING STATISTICS, TRENDS, AND FACTS

73% of school students say they feel they've been bullied in their lifetime. 73%

44% say they think they've been bullied in the last 30 days. 44%

Overall, 36.5% of people feel they've been cyberbullied in their lifetime. 36.5%

60% of teens say they have experienced some form of cyberbullying. 60%

70% of teens report that someone has spread rumors about them online. 70%

87% of young people have witnessed an instance of cyberbullying online. 87%

95% of teens actively use the internet, and regularly use social media. 95%

THRIVE my way

Insights on Cyberbullying

Digital Detox Bot

hello

Hello there! How can I assist you today?

What is up?

MODEL BUILDING

This section outlines the approach for model building. The process entailed leveraging a pre-trained base model, adapter-based fine-tuning using Parameter Efficient Fine-tuning (PEFT) configuration and incorporation of Reinforcement Learning from Human Feedback (RLHF) technique using Direct Preference Optimization(DPO) to create an effective agent for providing guidance for mental health improvement.

Additionally for text classification training a DistilBERT model and prompting the Mistral model was experimented with for classifying user inputs into social media addiction, social isolation and cyberbullying. Sentiment analysis was also performed to classify the text into three categories: mild, moderate and severe.

Key Components:

- **Base Model:** "mistralai/Mistral-7B-Instruct-v0.2" from Hugging Face. This pre-trained model, with a strong understanding of general language served as the foundation for the fine-tuning process and adapted for specific tasks.
- **Adapter:** "GRMenon/mental-health-mistral-7b-instructv0.2-finetuned-V2" - This adapter module contains task-specific knowledge related to mental health counseling conversations. It will be integrated with the base model to enhance its ability to handle mental health-related queries.
- **PEFT:** This configuration aimed to optimize memory usage and training efficiency during the fine-tuning process by focusing on just a subset of its parameters.
- **DPO**: Direct Preference Optimization technique which is a simplified RLHF was implemented which directly optimizes the LLM based on human preference data.
- **Preference data** in a specific format: prompt, chosen response (preferred), rejected response (not preferred).
- **Text Classification and Sentiment Analysis** using distilbert/distilbert-base-uncased,distilbert/distilbert-base-uncased-finetuned-sst-2-english and prompting the DPO trained Mistral were experimented with.
- **Guardrailing** the Mistral model using system prompts

Steps:

The following are the high level steps followed for model building:

1. **Load the Base Model:** "mistralai/Mistral-7B-Instruct-v0.2" with settings specifically chosen to reduce memory usage and optimize performance for the hardware it runs on.
2. **Create PEFT Model** -A PEFT model was created by combining the loaded base model with the PEFT configuration and the adapter "GRMenon/mental-health-mistral-7b-instructv0.2-finetuned-V2"

   This process effectively fine-tuned the base model's knowledge with the adapter's mental health-specific information.

3. **Implement Direct Process Optimization (DPO)**

○ TRL (Transformer Reinforcement Learning) library was used for DPO training

- ○ The model saved in the previous step was used for both the base model as well as reference model of DPO. The **reference model** ensures that whatever is generated does not deviate too much and continues to maintain generation diversity.
- ○ Since the models were trained using LoRa adapters, Peft's AutoPeftModelForCausalLM helper function was used to load them correctly.
- ○ The models are trained on the preference dataset containing the prompt-chosen-rejected format.
- ○ The model was configured using peft_config arguments to use a memory-efficient 4-bit format during training with the QLora method.
- ○ Finally, the DPO trained model was uploaded and shared on the **Hugging Face Hub.**

4. **Guard railing**

- ○ In order to assess Mistral's capacity to resist manipulation, adversarial prompts were employed to test its guard railing functionality.

5. **Text Classification and Sentiment Analysis**

- ○ The distilbert/distilbert-base-uncased model from Hugging Face was trained on the cleaned dataset to for classifying the user inputs into 3 classes: cyberbullying [0], social media addiction[1] and social isolation[2],
- ○ The fine tuned Mistral model was also prompt engineered to classify the user inputs into the 3 classes mentioned above.
- ○ Sentiment Analysis was performed by prompting the model to classify the text into three categories: mild, moderate and severe based on the severity of the text.

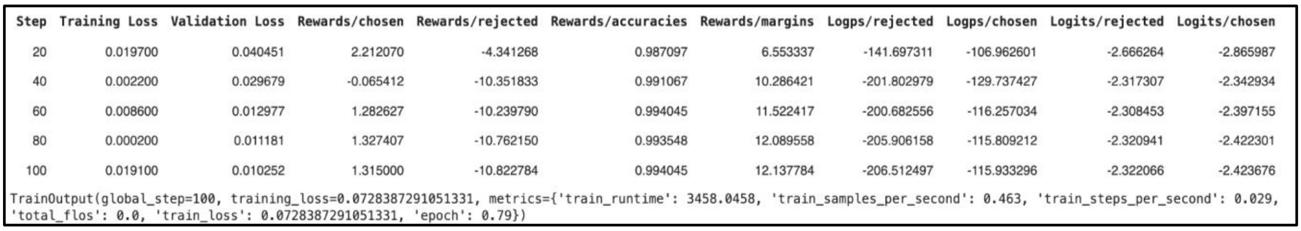
Evaluation Framework

An overview of the evaluation framework used to monitor the performance of the Direct Preference Optimization (DPO) trained model for mental health improvement is provided below:

- The performance of the training was monitored using a separate evaluation dataset and report metrics.
- These metrics were pushed to Weights and Biases (WandB) for ease of visualization and tracking
- The WandB logs contained the following reward metrics:
- ○ **rewards/chosen:** Measures the difference in preference between the model's chosen response and a reference model's response, scaled by a beta factor. Higher values indicate the model's chosen responses are more favorable.
- ○ **rewards/rejected:** Measures the difference in preference between the model's rejected responses and the reference model's response, scaled by beta. Lower values indicate the model is effectively rejecting less favorable responses.
- ○ **rewards/accuracies:** Represents the proportion of times the chosen response has a higher reward than the corresponding rejected response. Ideally, this metric should approach 1.0.
- ○ **rewards/margins:** Captures the average difference between the chosen response's reward and the corresponding rejected response's reward. A positive margin indicates the model is selecting preferable responses
- The evaluation goal was to:
- ○ **Maximize Chosen vs. Rejected Reward Margin:** This ensures the model prioritizes responses with higher preference compared to alternatives.
- ○ **Increase Accuracy of Chosen Responses:** The model should consistently select responses with higher rewards than rejected ones.

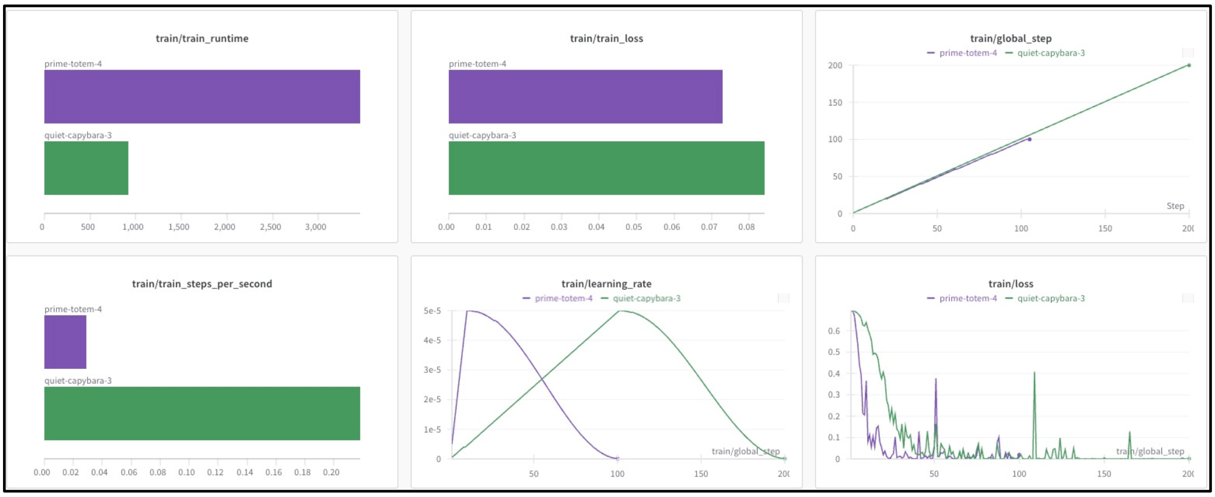
- The model performance was gauged by analyzing the calculated metrics to assess the model's ability to select supportive and appropriate responses in the mental health context.
  - A rising chosen vs. rejected reward margin indicates improvement in selecting better responses.
  - An accuracy metric approaching 1.0 suggests the model reliably chooses higher-rated responses.

A snapshot of the evaluation metrics and training is shown below:



Step	Training Loss	Validation Loss	Rewards/chosen	Rewards/rejected	Rewards/accuracies	Rewards/margins	Logps/rejected	Logps/chosen	Logits/rejected	Logits/chosen
20	0.019700	0.040451	2.212070	-4.341268	0.987097	6.553337	-141.697311	-106.962601	-2.666264	-2.865987
40	0.002200	0.029679	-0.065412	-10.351833	0.991067	10.286421	-201.802979	-129.737427	-2.317307	-2.342934
60	0.008600	0.012977	1.282627	-10.239790	0.994045	11.522417	-200.682556	-116.257034	-2.308453	-2.397155
80	0.000200	0.011181	1.327407	-10.762150	0.993548	12.089558	-205.906158	-115.809212	-2.320941	-2.422301
100	0.019100	0.010252	1.315000	-10.822784	0.994045	12.137784	-206.512497	-115.933296	-2.322066	-2.423676

```
TrainOutput(global_step=100, training_loss=0.0728387291051331, metrics={'train_runtime': 3458.0458, 'train_samples_per_second': 0.463, 'train_steps_per_second': 0.029,
'total_flos': 0.0, 'train_loss': 0.0728387291051331, 'epoch': 0.79})
```



- The guardrailing capability of the Mistral model was assessed by employing a collection of adversarial prompts.It was observed that the model refrained from responding to any of the queries within the tested set of adversarial prompts.

Additionally the output of text classification model like DistilBERT and DPO trained Mistral were evaluated on two aspects:

- **Performance Metrics:** These assessed how accurately the model categorizes text data and the sentiment analysis. Common metrics include accuracy, precision/recall, and F1 score, the results using DPO trained Mistral model are provided below:

```
              precision    recall  f1-score   support

    moderate       0.86      0.86      0.86         7
      severe       0.89      0.89      0.89         9

    accuracy                           0.88        16
   macro avg       0.87      0.87      0.87        16
weighted avg       0.88      0.88      0.88        16
```

- **Human Evaluation** While metrics provide a quantitative view, human evaluation was used to assess if the model's classifications aligned with real-world expectations. It was observed that the fined tuned Mistral model provided better results

Application Deployment

We opted to deploy our application on a Google Cloud Platform Virtual Machine, recognizing the requirements of utilizing a model with 7 billion parameters. The necessary GPU size for efficient inference was determined to be 24GB. Consequently, we configured the Virtual Machine with an L4 GPU, alongside 4 virtual CPUs and 16 GB of RAM to meet our computational needs.

The project is served through Streamlit, which allows for a clear division between backend and frontend functionalities. Specifically, the model_utils.py file manages model loading and executing inference functions. Meanwhile, the app.py file comprises the frontend development, including the design of a questionnaire, Exploratory Data Analysis (EDA) results, and an instance of the chat interface. User inputs like form responses and queries are directed to the model's inference function within the backend.

We implemented session state tracking to enhance user experience with faster loading times. This feature checks whether the GPU is already loaded into memory, thereby streamlining the inference process by eliminating the need for repeated model loading during refreshes. This approach not only accelerates the inference process but also conserves memory resources by avoiding redundant model loading within the same session.

In addition to these technical considerations, we took steps to ensure that our application is accessible to users globally. This involved modifying the firewall settings on the virtual machine to open port 8501, which our application uses. We specifically enabled both HTTP and HTTPS traffic to this port, ensuring secure and reliable access to our application from anywhere in the world.

CONCLUSION

In conclusion, our endeavor to address the multifaceted impacts of social media on human psychology has culminated in the creation of an interactive UI centered around an innovative LLM chatbot, employing Direct Preference Optimization. Recognizing the profound influence of social media on individuals, our approach combines personalized mental health support with community engagement to navigate the complexities of online interaction.

Our human-centric design ethos underscores the importance of holistic well-being in the digital age, aiming to mitigate negative consequences such as addiction, isolation, and cyberbullying. Leveraging a finely-tuned Mistral 7b LLM model trained on therapy-like counseling responses, our chatbot adeptly responds to a myriad of mental health issues arising from social media use.

Data-driven insights gleaned from user queries and sentiment analysis inform our chatbot's responses, ensuring relevance and appropriateness. The integration of a questionnaire within the UI facilitates targeted conversations, while future voice-to-text functionality intends to enhance user engagement.

Privacy remains paramount, with robust security measures safeguarding interactions and information. Deployed on the Google Cloud Platform, our chatbot stands ready to provide support to individuals worldwide, offering a personalized and secure resource in helping people free themselves from mental health issues caused by social media.

Limitations

This was a compressed 4 week project with about 50% of the team members becoming inactive over the span of that time. Challenges were encountered with sufficient GPU accessibility which caused delays to forward progress at the anticipated/preferred rate. Exploration of social media API usage for real-time UI integration into social media applications exposed various limitations such as pricing.

Future Directions

Recommendations for future work on this solution includes expansion to additional mental health subject areas outside of social media addiction, social isolation and cyberbullying. Future potential also includes real time integration for proactive intervention based on user preference settings.

RESOURCES

Links

- Omdena Challenge: Leveraging LLMs to Understand Global Mental Health Well-being & Fomo

Repository and models

- https://dagshub.com/Omdena/HyderabadIndiaChapter_MentalHealthWellbeingFomoSocialMedia/src/team-4
- https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
- https://huggingface.co/GRMenon/mental-health-mistral-7b-instructv0.2-finetuned-V2
- https://huggingface.co/gargvinayakk/dpo-mistralai-7b-mental-health
- GCP deployment: (IP address assigned to the VM is dynamic - each time the GCP VM is switched) As an example - here is one IP: http://34.143.189.165:8501/