**Turmerik Task Submission Report**

This document is a submission of the task for the ‘Machine Learning Engineering Intern’ position at Turmerik. This project aims to demonstrate my ability to scrape and analyze web data ethically, utilize sentiment analysis, and leverage AI to personalize communication. I was asked to focus on identifying potential participants for a clinical trial by analyzing sentiments expressed on Reddit.

The project has the following parts:

1. **Data Collection:** The data is being collected using the Reddit Data API. We are integrating this tool into the project using the PRAW framework in Python. The data is being collected in the following steps:

* First, we perform a keyword search for subreddits using the keywords: 'clinical trials,' 'clinical trial demographics,’ and 'clinical trial results to get a list of subreddits.
* Then, these subreddits are further filtered by performing cosine similarity on the description of each subreddit for each of these keywords. We do this to find the relevant subreddits from the search results. These results are exported in a CSV format.
* Then we extract the names of the subreddits with a similarity score of more than 0.65. Using these subreddit names, we scrape our data from reddit. The final data will be exported in a file called clinical\_trial\_data.csv

1. **Sentiment Analysis and User Segmentation:** Now after the raw data has been extracted, we will now perform sentiment analysis using Vader or OpenAI. The sentiment analysis results are structured in a JSON format. The sentiment analysis is being done in the following steps:

* First, we clean the scraped comments by removing URLs and non-alphabetic characters.
* Now, we group the data so that all comments written by an author in a subreddit are grouped together. The JSON structure follows the following structure:

username : {subreddit\_name :{ sentiment\_label:

average sentiment:

comments: }}

* The average sentiment is calculated using the average sentiment scores calculated by NLTK’s VADER model. If using OpenAI, the sentiment label that occurs in most comments written by a Redditor in a particular subreddit.
* After the data has been grouped, we perform sentiment analysis using Vader or Open AI. The results are exported in a json format where every author has it’s own value for sentiment in a subreddit.
* The JSON structure follows the following structure:

username : {subreddit\_name :{ sentiment\_label:

average sentiment:

comments: }}

1. Message Generation: The messages are generated using OpenAI’s Chat GPT-3.5 turbo model. To tune the LLM to perform the message generation task, the following prompt is given to it:

Create a personalized message for a Reddit user interested in {subreddit}, which is about '{subreddit\_description}.' They have a '{sentiment}' sentiment about participating in clinical trials. The message should encourage participation or provide informative feedback."},

* This prompt considers the following things while generating a personalized message for a user :

1. Subreddit Name
2. Subreddit Description
3. Sentiment Label

* Final Output:

"Manydanks": {

"ClinTexCTi": "hello there, it's great to see your interest in clintexcti and the world of clinical trials! it's wonderful to hear that you have a moderately positive sentiment about participating in clinical trials, as they play a crucial role in advancing medicine for the benefit of all.\n\nclintexcti's innovative approach with cti \u2013 clinical trials intelligence is truly groundbreaking. by lowering the costs of new medicines and speeding up their delivery to market, they are revolutionizing the way clinical trials are conducted. with their focus on making trials faster, more efficient, and more cost-effective, pharmaceutical companies can bring life-saving medications to those in need quicker than ever before.\n\nyour interest in this field is commendable, and by staying informed and engaged, you are contributing"

}

* The final output is a JSON file with the above structure.