Turmerik Take-home Assignment

Collecting Data

I utilized the Reddit PRAW for getting posts and comments from a subreddit. I focused on the Clinical trials subreddit.

Challenges:

1. Rate Limiting: I cannot constantly send a request as they have set the limit for time
2. The subreddit.get method limits to 1000 posts and comments which means I cannot go more
3. It was taking some time to backfill all the data

Some Solutions:

1. To get more posts and comments, I used hot, new, rising and other categories to get as many as I can
2. If needed, I added some sleep commands to delay each request

Privacy Concerns: I accepted all the privacy policy and agreement and do not have any personal identifiable information displayed

Sentiment Analysis

Now I do not have labelled data so supervised fine-tuning and training methods would had to be done on other datasets. Given time constraints, I relied on Vader.

Reasons for Using VADER

1. Designed for Social Media Text: VADER is specifically tailored for sentiment analysis of texts from social media platforms like Twitter and Reddit. It effectively handles slang, emoticons, acronyms, and shorthand, which are typical in social media text.
2. Lexicon-based Approach: VADER utilizes a human-curated sentiment lexicon that includes intensity measures for each word. This feature allows it to more accurately determine sentiments based on the lexical content of text.
3. VADER is designed to understand the impact of modifiers, such as intensifiers ("very good", “kind of bad”)
4. Real-time Analysis Capability: Due to its rule-based nature, VADER is fast and does not require any training data.

I also had to use NLTK and do some data preprocessing to remove some URL, special characters and things that could interfere with the sentiment.

I used the user's all the posts and comments to come with a post_sentiment, comment_sentiment and combined_sentiment. And then the score is between -1 and 1. I segregated from -0.1 to 0.1 to Neutral, 0.1+ to positive or receptive to clinical trials and the remaining are negative.

These are some of the word clouds and data distribution I had after running VADER and WordCloud.

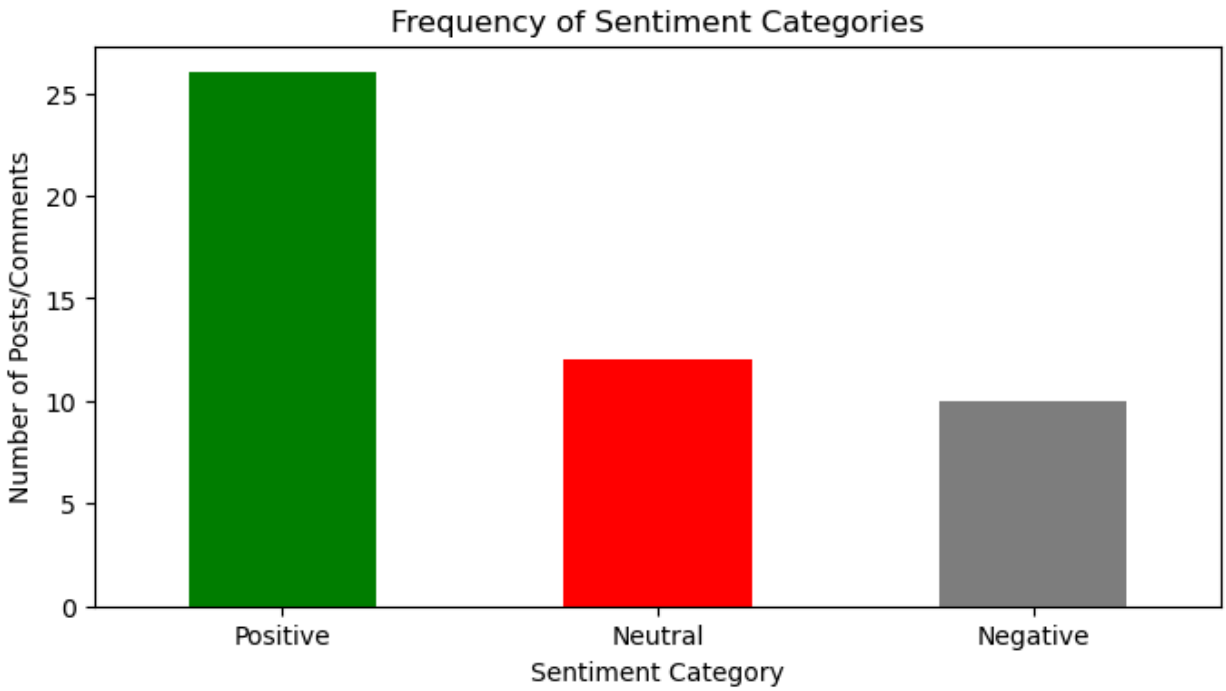
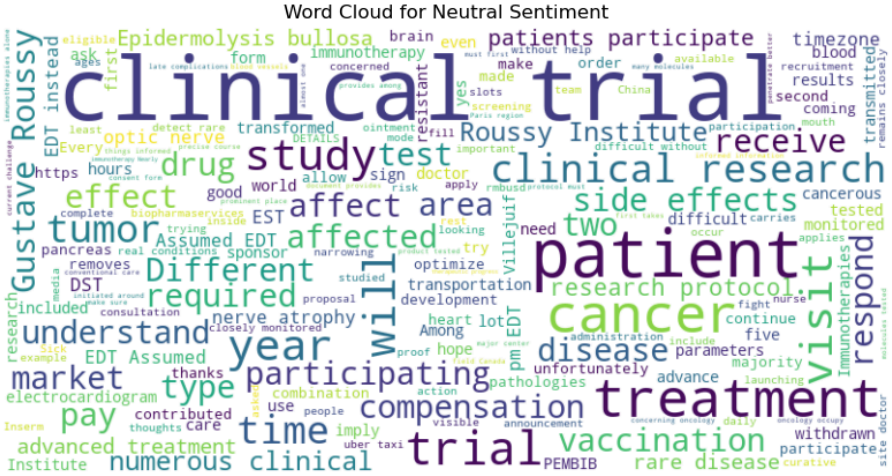


Figure 1: Sentiment Distribution

Word Cloud for Positive Sentiment

Word Cloud for Negative Sentiment

Word Cloud for Neutral Sentiment



The word clouds were interesting as for negative sentiments you see keywords like stress, trauma, disorder and others.

OpenAI Message Generation

I used gpt-3.5 turbo for message generation. I had a prompt which used the user's posts, comments, sentiment analysis and then generate a personalized message for them. These are some examples I got.

Example Output: Posts by the user:
- Research Study for Co-occurring Chronic Pain and Addiction - 100% online! The Biopsychosocial Pain Management Lab is looking for men and women ages 18-65 to participate in: **12 weeks of no-cost group tele-therapy and an additional 3 one-on-one sessions - 100% online.** Try a research tele-therapy treatment to reduce opioid use and pain - and be compensated for your time!* For more information, please contact Sarah with the Biopsychosocial Pain Management Lab of CU Denver at **(303) 578-8770** or [**stopstudy@hotmail.com**](mailto:stopstudy@hotmail.com). Eligibility includes a short phone screening prior to study participation. *Compensation is in the form of Target gift cards and is provided upon study completion. COMIRB# 17-1849 PI: Amy Wachholtz, PhD Research Study for Co-occurring Chronic Pain and Addiction - 100% online! The Biopsychosocial Pain Management Lab is looking for men and women ages 18-65 to participate in: **12 weeks of no-cost group tele-therapy and an additional 3 one-on-one sessions - 100% online.** Try a research tele-therapy treatment to reduce opioid use and pain - and be compensated for your time!* For more information, please contact Sarah with the Biopsychosocial Pain Management Lab of CU Denver at **(303) 578-8770** or [**stopstudy@hotmail.com**](mailto:stopstudy@hotmail.com). Eligibility includes a short phone screening prior to study participation. *Compensation is in the form of Target gift cards and is provided upon study completion. COMIRB# 17-1849 PI: Amy Wachholtz, PhD Research Study for Co-occurring Chronic Pain and Addiction - 100% online! The Biopsychosocial Pain Management Lab is looking for men and women ages 18-65 to participate in: **12 weeks of no-cost group tele-therapy and an additional 3 one-on-one sessions - 100% online.** Try a research tele-therapy treatment to reduce opioid use and pain - and be compensated for your time!* For more information, please contact Sarah with the Biopsychosocial Pain Management Lab of CU Denver at **(303) 578-8770** or [**stopstudy@hotmail.com**](mailto:stopstudy@hotmail.com). Eligibility includes a short phone screening prior to study participation. *Compensation is in the form of Target gift cards and is provided upon study completion. COMIRB# 17-1849 PI: Amy Wachholtz, PhD (Sentiment: -0.8521)

Comments by the user:
- (Comment Sentiment: 0.0)

Personalized Message:
Hey there! I noticed your interest in research studies for co-occurring chronic pain and addiction. Have you considered participating in the clinical trial offered by the Biopsychosocial Pain Management Lab at CU Denver? It's a great opportunity to try out a research tele-therapy treatment that could potentially help reduce opioid use and pain.

Not only will you be contributing to valuable research, but you'll also have the chance to receive compensation in the form of Target gift cards upon completion of the study. Plus, the convenience of it being 100% online makes it easily accessible for you.

If you're interested in learning more or getting involved, feel free to reach out to Sarah at (303) 578-8770 or stopstudy@hotmail.com. Your participation could make a real difference in advancing our understanding of chronic pain and addiction management. Let me know if you have any questions or need more information. Thanks for considering it!

So above was the post and if you see the personalized message, it includes specific details and something which the user can connect with more.

Challenges with OpenAI API

1. The text and post length may be too long, 60000 Tokens per minute for OpenAI
2. Context length window and LLM losing context in the middle
3. OpenAI API takes some time

Ways to Avoid this:

1. Summarized the text using a BART model when posts where too long
2. Counted how many characters in each request and added sleep when exceeded 60000

Some user_ids to experiment and see output: 7l9ag0z, 12qo9x, 84ech, etc. Can look at users.txt

Ethical considerations you took into account when designing and implementing your solution:

1. Accepted Reddit Guidelines
2. No personal identifiable information accessed from API
3. Followed API guidelines of Reddit and OpenAI

Improvements:

Some posts may have names or details, need to remove or delete them