Product Innovation Report

https://github.com/srivikas777/Sentiment-Analysis-on-Reddit-Data-using-ClinicalBERT.git

Brainstorm Process:

My brainstorming process began by selecting the disease on which clinical trials are being made diabetes. This helped me to narrow down my scope and focus properly. The main challenge was the lack of labeled data for sentiment analysis. So, I chose to use a pretrained model, ClinicalBERT. Even though I used ClinicalBERT, I still needed to train something on our relevant domain. I scraped some posts and comments with the keyword "Clinical Trials" and from some relevant subreddits. Then, I processed the data to remove null values, duplicates, and performed basic preprocessing and tokenization. Here arose another challenge: I didn’t even have labels here. So, I tried masking some words and training the model to predict those words, but the approach didn’t seem to work Ill. So, I used vaderSentiment to generate labels for clinical data. As the data was already relevant to clinical trials, I thought labeling based on tone and words would be more suitable than labeling diabetes data as it is not that relevant to our domain. Then, I used labeled data to fine-tune the ClinicalBERT model. Due to limited time, I only tried a few parameters and chose the best one out of them. Now, I saved the fine-tuned model and scraped data from the diabetes subreddit. I performed similar preprocessing steps here too and used the fine-tuned model to predict sentiments towards clinical trials. Then, I separated positive sentiment data and used those texts to generate personalized messages using the OpenAI API.

Product Requirements:

- **Sentiment Analysis Functionality:**
    - The primary objective was to develop robust sentiment analysis capabilities focusing on Reddit posts and comments related to diabetes and clinical trials.
- **Personalized Message Generation:**
    - Once I know the mood and interest of the user towards clinical trials via sentiment analysis, a key feature was the generation of personalized messages based on sentiment analysis results. These messages aimed to encourage participation in clinical trials.
- **Privacy and Security:**
    - Ensuring data privacy and compliance with Reddit's policies Ire paramount. I took careful measures to safeguard user information and respect platform guidelines.

Minimum Viable Product (MVP) Plan:

- **Data Scraping Module:**
    - Implementing data scraping from Reddit using keywords or subreddit names enabling us to access as much relevant data as possible. Then, processing it.
- **Sentiment Analysis Module:**

  - Implementation of sentiment analysis using the ClinicalBERT model was the cornerstone of our MVP. I fine-tuned the model using data collected from Reddit, specifically focusing on posts mentioning "Clinical Trials." A few steps were taken according to the situation to cross the challenges.

- **Message Generation Module:**

  - I developed a module to generate personalized messages based on sentiment analysis results. Leveraging the OpenAI API, I crafted messages aimed at encouraging participation in clinical trials.

Future Experimentation Plan:

- **Goal: Evaluate the effectiveness of personalized messages.**

  - **Experiment:** Conduct A/B testing with different message variations.
  - **Metric:** Measure the increase in engagement with clinical trial-related posts.

- **Goal: Assess sentiment analysis accuracy and user satisfaction.**

  - **Experiment:** Solicit user feedback through surveys post-platform usage.
  - **Metric:** Analyze user satisfaction scores and sentiment analysis performance metrics.

Long-term Evolution:

- Getting more structured and labeled data via trustworthy sources or starting self-data collection and manual labeling.

- I envision partnering with healthcare providers to provide insights into community sentiment for better patient care and treatment strategies. And, to get access to large customer data and the ability to collect data.

- Focusing more on fine-tuning and preprocessing maintaining the quality of data and respectively sentiment analysis.