**Sentiment Analysis on Diabetes Reddit using ClinicalBERT**

**Introduction**

The aim of this project is to perform sentiment analysis on posts and comments related to diabetes on Reddit. Given the lack of labeled data for sentiment analysis in this domain, we utilized a pre-trained model, ClinicalBERT, and fine-tuned it using data collected from Reddit, particularly focusing on posts and comments mentioning "Clinical Trials." The project also involved generating personalized messages based on sentiment analysis results.

**Setup Instructions:**

* Generate Reddit API Credentials:
  1. Obtain API credentials from Reddit and replace them in the code where specified.
* Replace OpenAI API Key and Organization:
  1. Obtain an API key from OpenAI and specify your organization details in the code where required.
* Accessing Datasets and Pre-trained Models:
  1. All datasets of scraped data are included in the provided folder.
  2. The fine-tuned model for sentiment analysis is also included in the folder and can be directly used.
  3. If you choose to run the fine-tuning code again, ensure to replace the path to the new saved model in the sentiment analysis code for importing the model.
* Running the Code:
  1. After completing the above steps, you can simply run the code.
  2. Ensure all necessary dependencies are installed and run all cells in the provided notebooks in a regular manner.

**Methodology:**

* Data Collection:
  + We collected posts and comments from relevant subreddits using keywords such as "Clinical Trials" and other diabetes-related terms.
  + The collected data was then preprocessed to remove null values, duplicates, and underwent basic text preprocessing and tokenization.
* Labeling Data:
  + Since we lacked labeled data for sentiment analysis, we used the VADER sentiment analysis tool to generate labels for the clinical data based on the tone and words used.
  + VADER provided sentiment scores for each text, which were used as labels for training our sentiment analysis model.
* Model Fine-Tuning:
  + We fine-tuned the pre-trained ClinicalBERT model using the labeled data generated from the VADER sentiment analysis.
  + Due to time constraints, a limited set of hyperparameters were explored, and the best-performing model was selected.
* Sentiment Analysis on Diabetes Reddit:
  + After fine-tuning the model, we applied it to the collected data from the diabetes subreddit.
  + Similar preprocessing steps were applied to clean and prepare the data for sentiment analysis.
* Message Generation:
  + From the posts and comments with positive sentiment towards clinical trials, personalized messages were generated using the OpenAI API.
  + These messages aimed to provide encouragement and support to individuals interested in participating in clinical trials for diabetes.

**Challenges:**

1. Unlabeled Data: The lack of labeled data for sentiment analysis posed a significant challenge. We addressed this by utilizing VADER sentiment analysis for labeling.
2. Model Training Time: Training models on our machine proved to be time-consuming due to the complexity of the ClinicalBERT model and the large dataset.
3. Interdependencies: The project involved numerous interdependencies between data collection, preprocessing, model training, and message generation, requiring extensive brainstorming and trial-and-error.

**Examples of Data Collected, Analysis Performed, and Messages Generated:**

* Data Collected:
  + Sample Post from Clinical Trials Subreddit:

A close-up of a document

Description automatically generated

* + Word Cloud

A close up of words

Description automatically generated

* Analysis Performed:
  + Sentiment Analysis Labeling:

A screenshot of a computer

Description automatically generated

* Messages Generated:
  + Personalized Message for Clinical Trials Enthusiast:

A screenshot of a computer

Description automatically generated

**Ethical Considerations:**

1. Privacy and Data Usage: Care was taken to ensure that the collected data did not include personally identifiable information (PII) of Reddit users, and Reddit's terms of service and API usage policies were respected.

2. Bias Mitigation: While Reddit data was used for sentiment analysis, potential biases in the platform's user demographics were acknowledged. Steps were taken to mitigate bias in sentiment analysis algorithms to ensure fair representation of user opinions (Planned to handle but due to very limited data I was unable to do this).

3. Consequences of Messaging: The personalized messages generated were crafted to be accurate, respectful, and sensitive to potential vulnerabilities. Consideration was given to the potential impact of the messages on individuals.

**Conclusion:**

In conclusion, this project demonstrated the application of sentiment analysis techniques, utilizing a pre-trained model and fine-tuning it for a specific domain without labeled data. Despite challenges such as data collection, model training time, and ethical considerations, the project successfully generated insights into sentiment towards clinical trials in the diabetes community on Reddit and provided personalized messages to encourage participation in clinical trials.