Sentiment Analysis on Diabetes Reddit using ClinicalBERT

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

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:
    a. Obtain API credentials from Reddit and replace them in the code where specified.
- Replace OpenAI API Key and Organization:
    b. Obtain an API key from OpenAI and specify your organization details in the code where required.
- Accessing Datasets and Pre-trained Models:
    c. All datasets of scraped data are included in the provided folder.
    d. The fine-tuned model for sentiment analysis is also included in the folder and can be directly used.
    e. 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:
    f. After completing the above steps, you can simply run the code.
    g. 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.

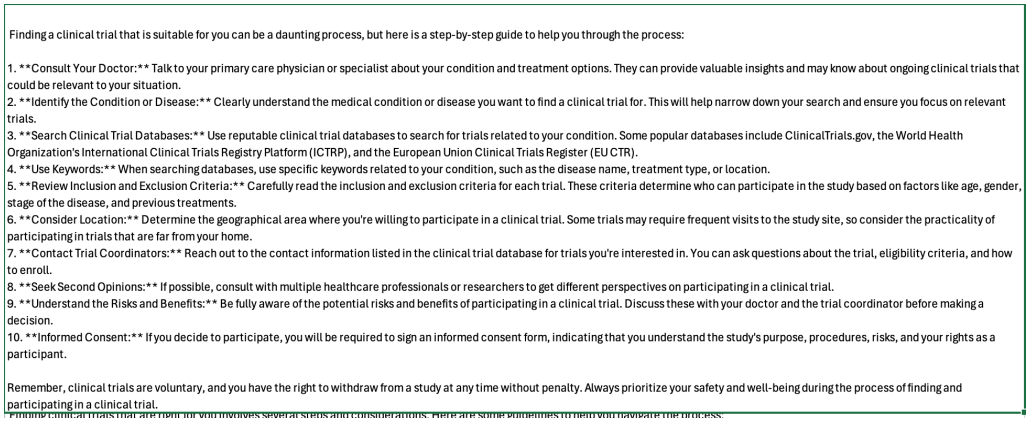
    - Due to time constraints, a limited set of hyperparameters were explored, and the best-performing model was selected.
- Sentiment Analysis on Diabetes Reddit:
    - After fine-tuning the model, we applied it to the collected data from the diabetes subreddit.
    - Similar preprocessing steps were applied 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:



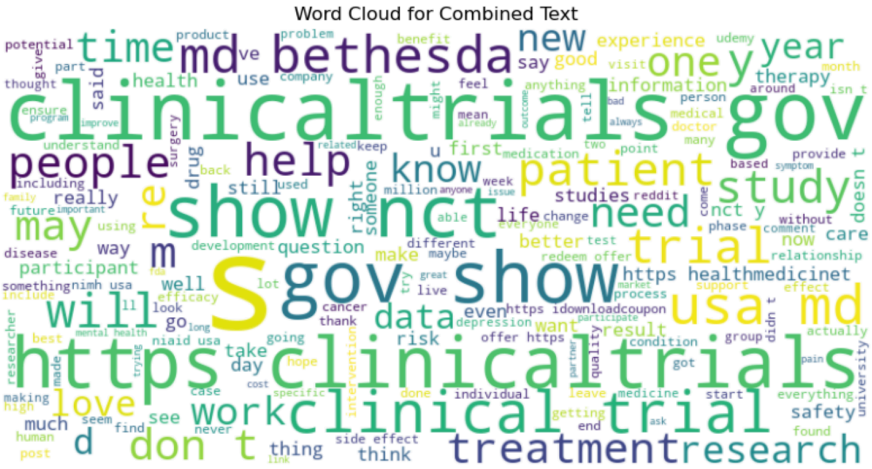
Finding a clinical trial that is suitable for you can be a daunting process, but here is a step-by-step guide to help you through the process:

1. **Consult Your Doctor:** Talk to your primary care physician or specialist about your condition and treatment options. They can provide valuable insights and may know about ongoing clinical trials that could be relevant to your situation.
2. **Identify the Condition or Disease:** Clearly understand the medical condition or disease you want to find a clinical trial for. This will help narrow down your search and ensure you focus on relevant trials.
3. **Search Clinical Trial Databases:** Use reputable clinical trial databases to search for trials related to your condition. Some popular databases include ClinicalTrials.gov, the World Health Organization's International Clinical Trials Registry Platform (ICTRP), and the European Union Clinical Trials Register (EU CTR).
4. **Use Keywords:** When searching databases, use specific keywords related to your condition, such as the disease name, treatment type, or location.
5. **Review Inclusion and Exclusion Criteria:** Carefully read the inclusion and exclusion criteria for each trial. These criteria determine who can participate in the study based on factors like age, gender, stage of the disease, and previous treatments.
6. **Consider Location:** Determine the geographical area where you're willing to participate in a clinical trial. Some trials may require frequent visits to the study site, so consider the practicality of participating in trials that are far from your home.
7. **Contact Trial Coordinators:** Reach out to the contact information listed in the clinical trial database for trials you're interested in. You can ask questions about the trial, eligibility criteria, and how to enroll.
8. **Seek Second Opinions:** If possible, consult with multiple healthcare professionals or researchers to get different perspectives on participating in a clinical trial.
9. **Understand the Risks and Benefits:** Be fully aware of the potential risks and benefits of participating in a clinical trial. Discuss these with your doctor and the trial coordinator before making a decision.
10. **Informed Consent:** If you decide to participate, you will be required to sign an informed consent form, indicating that you understand the study's purpose, procedures, risks, and your rights as a participant.

Remember, clinical trials are voluntary, and you have the right to withdraw from a study at any time without penalty. Always prioritize your safety and well-being during the process of finding and participating in a clinical trial.

Finding clinical trials that are right for you involves several steps and considerations. Here are some guidelines to help you navigate the process:

- Word Cloud



- Analysis Performed:
  - Sentiment Analysis Labeling:

text	opinion	sentiment	sentiment_probabilities
the future of clinical trials is bright with many exciting trends on the horizon here are sc	positive	negative	{'positive': 0.049453236162662506, 'negative': 0.9505468010902405}
protecting your privacy when participating in a clinical trial is essential to ensure your pi	positive	negative	{'positive': 0.046325597912073135, 'negative': 0.9536744356155396}
there are many places where you can find clinical trials here are a few of the most popul	positive	negative	{'positive': 0.04482852295041084, 'negative': 0.9551714658737183}
discussing the possibility of participating in a clinical trial with your doctor is an import	positive	negative	{'positive': 0.04760678857564926, 'negative': 0.9523932337760925}
participating in a clinical trial can offer benefits but it also involves certain risks that pa	positive	negative	{'positive': 0.057070884853601456, 'negative': 0.9429291486740112}
participating in a clinical trial can offer several benefits for both individuals and the broa	positive	negative	{'positive': 0.057300787419080734, 'negative': 0.9426991939544678}
clinical trials are research studies conducted on humans to evaluate the safety and effe	positive	negative	{'positive': 0.050124939531087875, 'negative': 0.9498750567436218}
finding a clinical trial that is suitable for you can be a daunting process but here is a ste	positive	negative	{'positive': 0.047167353332042694, 'negative': 0.9528326392173767}
finding clinical trials that are right for you involves several steps and considerations he	positive	negative	{'positive': 0.04957413673400879, 'negative': 0.9504258632659912}
finding clinical trials for a specific medical condition involves several steps and resourc	positive	negative	{'positive': 0.04629680514335632, 'negative': 0.9537031650543213}
if you are interested in participating in a clinical trial here are some general steps you ca	positive	negative	{'positive': 0.04696998745203018, 'negative': 0.953029990196228}
trial title a phase randomized double blind placebo controlled trial to evaluate the ef	negative	negative	{'positive': 0.09159453213214874, 'negative': 0.9084054231643677}
trial title a phase randomized double blind placebo controlled trial to evaluate the ef	positive	negative	{'positive': 0.08384300768375397, 'negative': 0.9161569476127625}
the novartis nct trial is a phase randomized double blind placebo controlled trial e	positive	negative	{'positive': 0.14163342118263245, 'negative': 0.8583666086196899}
trial title efficacy and safety of bnt b mrna covid vaccine trial status completed	positive	negative	{'positive': 0.3181397318840027, 'negative': 0.6818602681159973}
trial name a phase randomized double blind placebo controlled study to evaluate th	positive	negative	{'positive': 0.0566755011677742, 'negative': 0.9433244466781616}

- Messages Generated:
  - Personalized Message for Clinical Trials Enthusiast:

have you been experiencing illness pain for years but don t have a diagnosis have you been feeling lonely or isolated i want to hear about your experiences i m looking for people aged with undiagnosed physical symptoms for or more years who experience loneliness isolation to take part in my study you will be asked to complete a confidential online questionnaire about your experiences if you re interested please follow this link for more info and to get to the study https uwe eu qualtrics com jfe form sv vw lpskufvwh https uwe eu qualtrics com jfe form sv vw lpskufvwh thanks for participating	Hello, I acknowledge your journey of living with undiagnosed physical symptoms and the challenges it brings. Your courage to share your experiences is truly commendable. I am conducting a study focusing on individuals aged [age range] who have faced similar struggles for [number of years] or more, dealing with loneliness and isolation. Your participation could provide valuable insights into this complex issue. If you are interested in contributing to this study, please click on the following link for more details and to participate: [insert study link]. Thank you for considering joining this important research effort. Best regards, [Your Name]

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.