Title: Health and Wellness Survey

Table of Content:

Abstract:

Introduction:

Data:

In the study, two phases were employed in the data collection method to fulfill the requirement presented by CogniXR. The two phases are shown in the flow chart below:

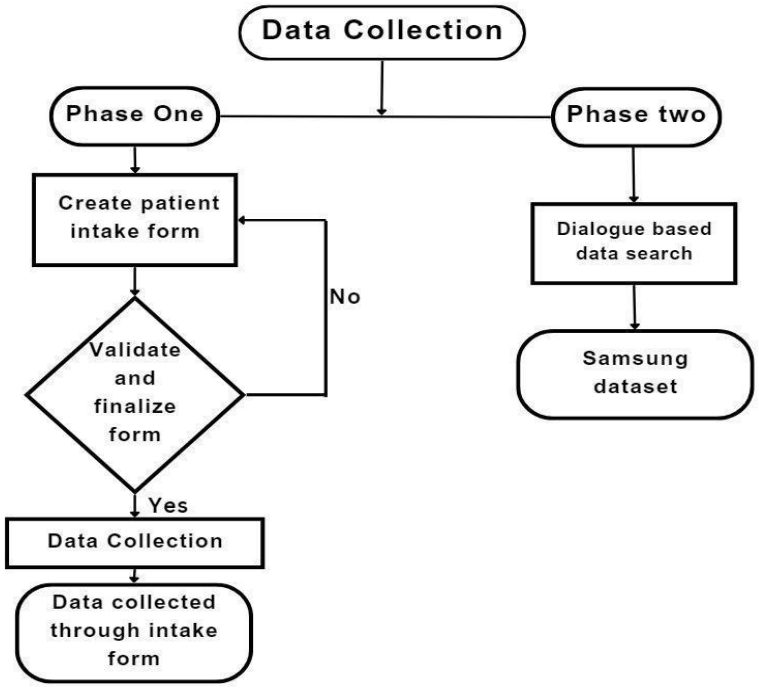

Figure 1: Data Collection

Data Collection Method:

Initially, an intake form was designed with questions that were focused on four mental health issues: Anxiety, Trauma, Substance Abuse, and PTSD. These questions were based on the initial screening questions asked by the therapist to the patient. This questionnaire was developed with multiple revisions and the help of CogniXR. The questionnaire consisted of demographic questions, and the impacts of their mental health condition on their sleep patterns, relationships, and daily lives. The form was created in Microsoft Forms maintaining the ethical aspects of a survey. After the finalization of the questionnaire by CogniXR, the study further proceeded to data collection.

The initial phase consisted of the collection of the data through the development of a set of questionnaires on Microsoft Forms focussing on four Mental health issues portfolio: Anxiety, PTSD, Trauma, and Substance Abuse with the help of a CogniXR therapist. Through multiple revisions of the questions and maintaining ethical integrity throughout the questionnaire was finalized.

After the questionnaire was finalized, we started collecting the data by distributing the Microsoft link to the participants. The targeted participants were the ones who were facing

any one of the four mental health conditions Anxiety, PTSD, Trauma, and Substance Abuse. This questionnaire was based on the therapist's initial screening of a patient.

The questionnaire focuses on the effect of their mental health condition on their daily life, sleep pattern, and relationship with the people around them. This also enquires about their current diagnosis, previous history, and experience.

The second phase required us to use the Samsung dataset which has 16k records of messenger-like conversation dialogues and their summary in the English language to create the summary for different questionnaires that CogniXR has been using for the employee's Mental Health inquiry. The SAMSum dataset was prepared by the Samsung R&D Institute Poland and distributed for research purposes under a non-commercial license: BY-NC-ND 4.0 [1]

Data Description:

The first dataset which was collected from the Microsoft form consists of 71 rows and 93 columns which consist of a diversity of people spread among the four personas Anxiety, PTSD, Substance Abuse, and Trauma.

The second dataset i.e. the SAMSum dataset contains 16369 conversations that were created by linguists who are fluent in English, they created real-life conversations reflecting their daily life. The style consists of informal, semi-formal, and formal type of conversations that makes it diversified on the basic conversation patterns. The entered data also consists of Slang words, emoticons, and typos which were annotated with summaries. About 75% of the dialogues in the dataset consist of a conversation between two participants and the rest 25% consists of more than two participants. The dataset consists of three columns and 16369 rows. The columns are:

- Dialogue: text of the dialogue.
- Summary: human-written summary of the dialogue.
- Id: unique id of an example.

Table 1: Samsung Dataset

Training Samples	**Testing Samples**	**Validation Samples**
14,732	819	818

Method of Analysis:

Fine-tuning BART for questionnaire summarization

The Bart model was proposed in BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension by Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer on 29 Oct 2019.[2]

Pre-trained Model - BART checkpoint on CNN-DM dataset.[4]

Fine-tuning Process:

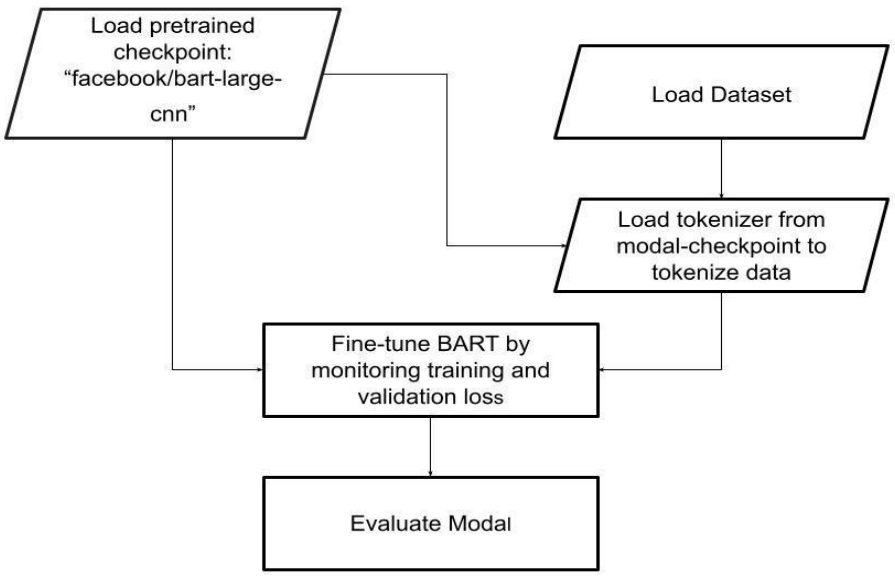

Figure 2: Fine-tuning Process

Evaluation Metrics

Loss: Sum of errors in each example of the evaluation set

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation:

Original Paper: Lin, Chin-Yew. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.[3]

(Plots for the evaluation set)

Evaluation of Validation Set

Table 2: Evaluation of validation set

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	RougeL	RougeLsum	Gen Len
0	1.358300	1.397652	41.83	21.73	32.47	38.84	59.60
1	1.055400	1.367195	41.73	21.53	32.27	38.64	60.12
2	0.837700	1.437623	41.53	21.12	31.89	38.46	60.16

Evaluation of Training Set

Table 3: Evaluation of Training set

Validation Loss	Rouge1	Rouge2	RougeL	RougeLsum	Gen Len
1.437623	41.537	21.121	31.89	38.465	60.16

Results:

Discussion:

Conclusion:

Acknowledgment:

References:

1. samsum · Datasets at Hugging Face. (2023). Retrieved 22 March 2023, from https://huggingface.co/datasets/samsum
2. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., & Levy, O. et al. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Retrieved 18 March 2023, from https://arxiv.org/abs/1910.13461
3. (2023). Retrieved 22 March 2023, from https://aclanthology.org/W04-1013.pdf
4. Facebook/Bart-Large-CNN · hugging face. (n.d.). Retrieved March 22, 2023, from https://huggingface.co/facebook/bart-large-cnn
5.
6.