

# Health and Wellness Survey

Post-Graduate CAPSTONE Project Report

In the partial fulfilment of requirement for the  
Data Analytics For Business (DAB)



**Supervised by:**

Mr. Umair Durrani

Professor, St. Clair College Center for the Arts

**Submitted by:**

Shailsuman Singh (0786915)

Naga Vijaya Pettichetti (0795708)

Atish Adhikari (0766698)

Aabriti Karki (0781762)

Somya Agrawal (0792049)

## Table of Content:

Abstract:.....	2
Introduction: .....	3
Data: .....	4
Data Collection Method: .....	4
Data Description: .....	5
Methodology: .....	7
Clustering: .....	7
Mathematical model: .....	9
Fine-tuning BART for questionnaire summarization.....	10
Results:.....	13
Conclusions and Recommendations .....	15
Acknowledgment:.....	17
References: .....	18

## Abstract:

According to the World Health Organization (WHO) mental health as “a state of well-being in which the individual realizes his or her own abilities, can cope with the stresses of life and can work productively and fruitfully and is able to make contribution to his or her community”. Mental health stability is very essential for a person, but the resources allocated by any country are insufficient which has led to a gap of more than 70% (*Mental Health - PAHO/WHO | Pan American Health Organization*, n.d.). According to WHO depression is the most common mental disorder in the world, CogniXR is a platform which believes in quality access of mental health services to everyone.

The main purpose of the project is to fulfil the requirement given by CogniXR which is show the severity of the patient such that the therapist can filter the patient in a data-driven way. Also, create a summary report for the questionnaire provided by the therapist. We create a questionnaire that uses scoring logic to display the severity level in an automated email to the therapist. And trained the BART model such that it creates a summary report of any questionnaire that is filled by the respondent which is then forwarded to the therapist.

# Introduction:

CogniXR has created a platform to increase the accessibility of mental healthcare solutions so that medical professionals and wellness initiatives may assist more people in achieving more balanced recovery. One of the most pressing challenges of our time is mental health. 10% of the worldwide disease burden and 25.1% of non-fatal disease burden are accounted for by mental, neurological, and substance use disorders (*10 Facts on Mental Health*, n.d.). Since there are not enough therapists to patients, these easily accessible platforms will be crucial in raising the ratio.

In this project, we have developed a system that will allow us to shorten the time that passes when a patient sees a mental health practitioner for the first time. Any patient who sees a therapist for the first time will fill out an intake form, which asks for information about their personal lives as well as information about their treatment, what kind of treatment they are seeking.

The therapist reviews the information the patient gave in the intake form and attempts to determine the best course of action for that patient after receiving it. With the aid of the knowledge, we have learned from Data Analytics for Business starting in the year 2022, we are attempting to make that process easier.

This project may assist a professional therapist in terms of primary analysis and thinking procedure to anticipate the treatment strategies. Help in lowering the cost of treatment by automating the process. So, for this we need to create a data driven approach to determine the severity of treatment and a competitive edge to reduce time consumption for filtration of patients.

The two objectives of this project

1. Prepare a psychometric questionnaire through which severity can be calculated and automated summary email would be sent to the therapist.
2. Develop an application which can be used to create any questionnaire and generate a summary report which would be sent to the therapist.

## Data:

In the study, two phases were implemented in the data collection method to fulfil the requirement presented by CogniXR. The two phases are shown in the flowchart below:

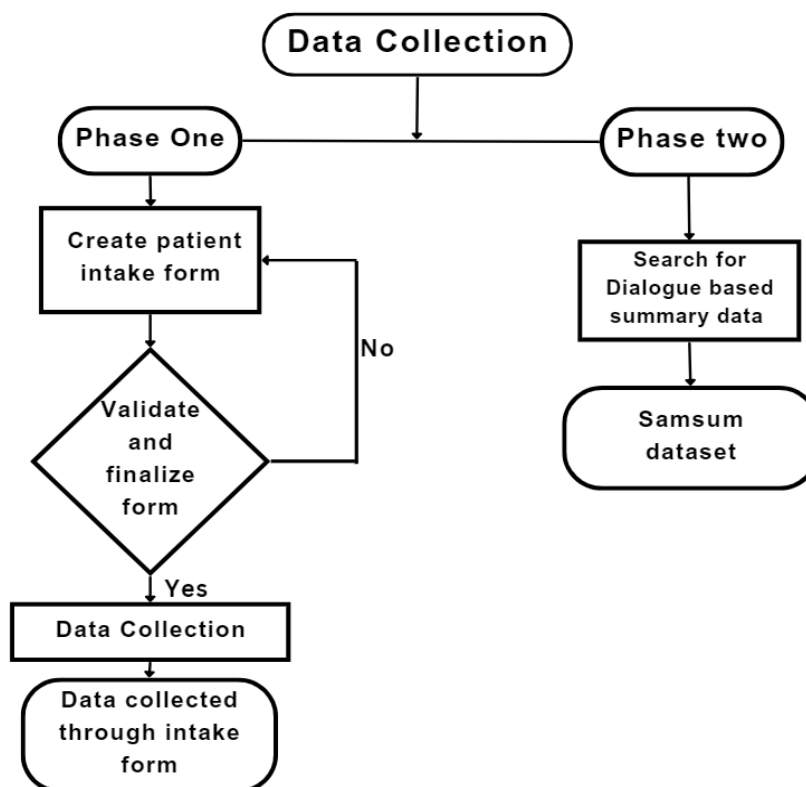


Figure 1: Data Collection Flowchart

## 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 finalised, 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.

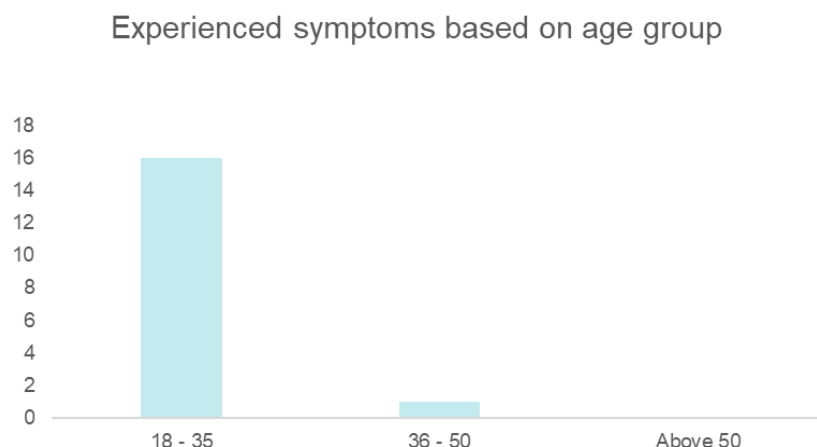
The questionnaire focuses on the effect of their mental health condition on their daily life, sleep pattern, and relationship with the people around them. This also enquires about their current diagnosis, previous history, and experience.

The second phase required us to use the SAMSum dataset to create the summary for different questionnaires that CogniXR has been using. The SAMSum dataset was prepared by the Samsung R&D Institute Poland and has 16k records of messenger-like conversation dialogues and their concise summary in the English language(*Samsum · Datasets at Hugging Face*, 2022).

## Data Description:

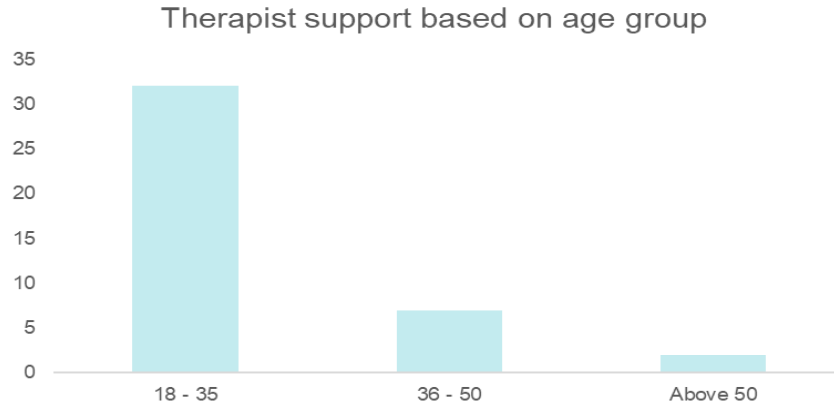
The first dataset which was collected from the Microsoft form consists of 71 rows and 93 columns which consist of a diversity of people spread among the four personas Anxiety, PTSD, Substance Abuse, and Trauma.

The following figures described the kind of data that was collected using the questionnaire created.



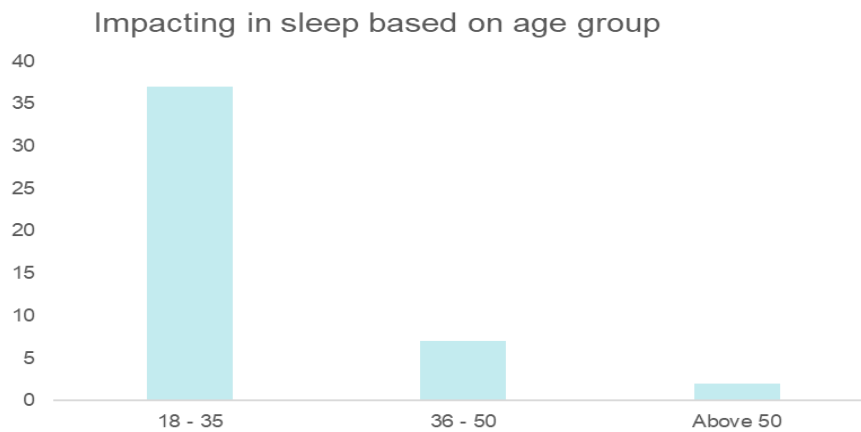
*Figure 2: Experienced symptoms based on age group.*

We had asked the potential patients to tell us if they had experienced any symptoms of their condition recently, those who replied yes are plotted against the age group in the above graph. As we had conducted the survey in St. Clair College and among similar age group to us the major number of participants are from 18 to 35 years of age group.



*Figure 3: Therapist support based on age group.*

In the above figure we had asked the participants if the therapist was able to make good impact in their mental health condition and most of them replied that they had good effect on them.



*Figure 4: Impacting in sleep based on age group.*

As we had asked the participants if they experienced any changes in their sleep pattern, above figure shows that most of the participants had impacts in their sleep as a result of their mental health condition.

The second dataset i.e., the SAMSum dataset contains 16,369 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. (Samsum · Datasets at Hugging Face, 2022). The columns are:

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

Table 1: SAMSum dataset

Training Samples	Testing Samples	Validation Samples
14,732	819	818

We had used 14732 samples as training set, 819 samples as the testing set and 818 samples as validation set for our model.

## Methodology:

The first phase of the project aim was to find severity based on the questions (first dataset) and to provide summarization of the questionnaire data in email. The severity would denote the urgency of the patient where Red (Highly severe), Yellow (Moderate severity), Green (Low severity). Using this severity would help the mental health providers to address the patient according to the urgency of the mental health patient. Along with this severity report would be sent to the provider once the patient fills out a form.

## Clustering:

The first approach we used was the clustering which would group the data in three sets. We used different clustering methods to group the data i.e., DBscan, K-means, K-Mode, Hierarchical-Agglomerative. After trying these clustering methods, we went with the K-Means clustering since the silhouette score was high for K-means clustering.

The diagram below shows that we were able to make three clusters. In the diagrams below we have used Principal Component Analysis (PCA) and t-distributed stochastic neighbourhood embedding (t-SNE) to visualize the data and clusters in two dimensions.



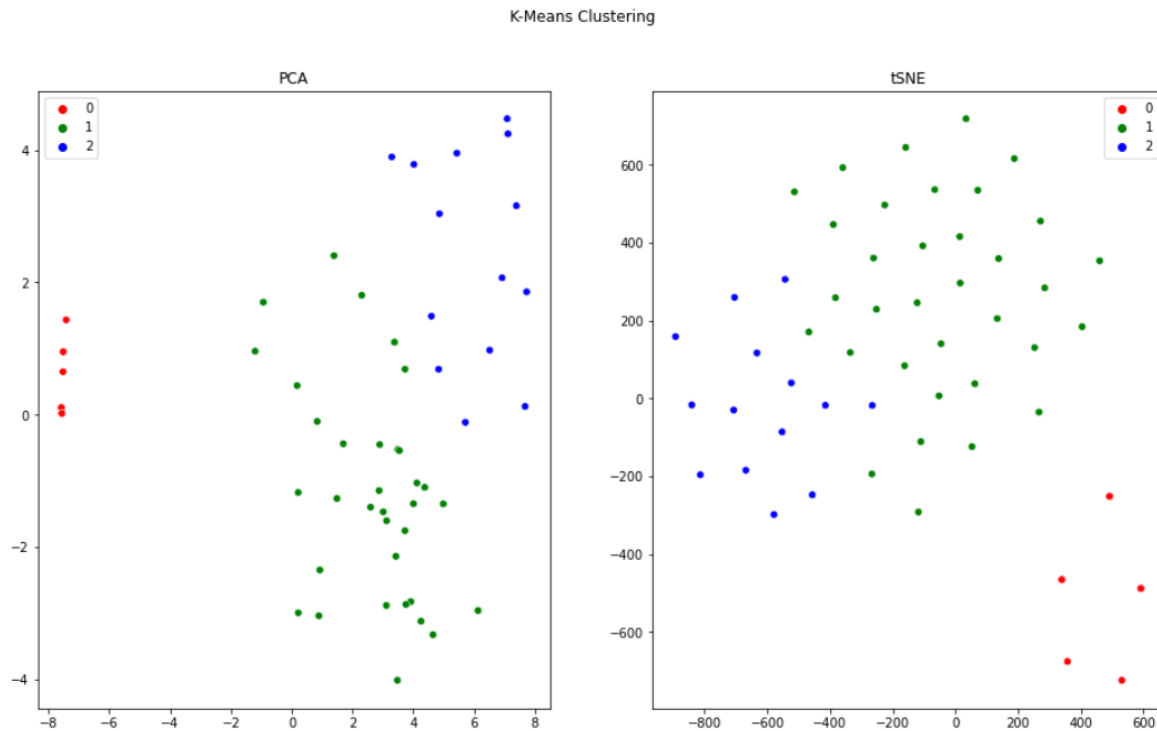


Figure 5: Two-dimension representation of clusters formed using K- Means

Although we were able to form clusters, we still couldn't use this method. The three clusters that were formed showed mixed data, and no conclusions could be made for each of the clusters.

The screenshot below shows the two of the clusters that were formed and looking at one of the attributes in the clusters we can see that there is mixed data. The cluster 1 has Moderately, A little bit, Quite a bit, Extremely and so does cluster 2.

Table 2: Table showing mixed data in the same cluster

therapist_support_anxiety	daily_affect_anxiety	goals_anxiety	expectations_anxiety	relationship_changes_anxiety	sleep_impact_anxiety	sleep_patterns_anxiety	n
NaN	NaN	NaN	NaN	NaN	NaN	NaN	
Adequate	Prone to more isolation	To help me relax;To have better relationships;...	To feel relax;To learn coping strategies;To fe...	No not at all	Quite a bit	Quite a bit	
Adequate	Prone to more isolation	To help me relax;To smile again;To have better...	To feel relax;To learn coping strategies;To ha...	No in other environments	Quite a bit	Quite a bit	
NaN	NaN	NaN	NaN	NaN	NaN	NaN	
NaN	Prone to more anger	To help me relax;To smile again;To have better...	To feel relax;To learn coping strategies;To ha...	Yes mostly with my friends	Moderately	Moderately	
Haven't been to therapy	Prone to more grief	To help me relax;To smile again;To have better...	To feel relax;To learn coping strategies;To ha...	Yes mostly with my friends	Extremely	Quite a bit	

This attribute is just an example of the heterogeneous data that was seen. We have seen the different variations of data in other attributes as well. A label could not be assigned to these clusters hence we couldn't use this method.

## Mathematical model:

To calculate the severity of the patient we went with the scoring logic and assessed the type of therapy the patient will be requiring. Similar technique has been used by National Center of PTSD, USA for calculating the severity of PTSD in PCL-5 questionnaire (*PTSD Checklist for DSM-5 (PCL-5) - PTSD: National Center for PTSD*, n.d.). CES-D questionnaire (*Center for Epidemiological Studies Depression (CESD)*, n.d.) and Hamilton-D questionnaire for depression also use similar techniques (*HAMILTON-DEPRESSION.Pdf*, n.d.). PWSQ questionnaire for worry also uses similar scoring for answer choices and gives the severity based on total score (*Penn State Worry Questionnaire (PSWQ) — Nathan Kline Institute - Rockland Sample Documentation*, n.d.). Each answer choice is given a score. Scores are added together to determine severity. Words like suicidal, suicide, murder, 911, urgent directly flag response as severe.

Here is the example of the scoring logic. We have used a score of 0-3 range. The scores for user selected answers are added together to give a total score that determines severity.

Q. When was the last time you have experienced symptoms of Anxiety?

- ☒ a. < 1 month
- b. < 3 months
- c. < 6 months
- d. < 1 year
- e. > 1 year

Q. Has it affected your sleep pattern?

- a. Not at all
- b. A little bit
- c. Moderately
- d. Quite a bit
- ☒ e. Extremely

Q. What are your expectation from that treatment?

- ☒ a. To feel relaxed
- b. To learn coping strategies
- ☒ c. To have better relationships
- d. To feel better at work or school
- e. To be happy

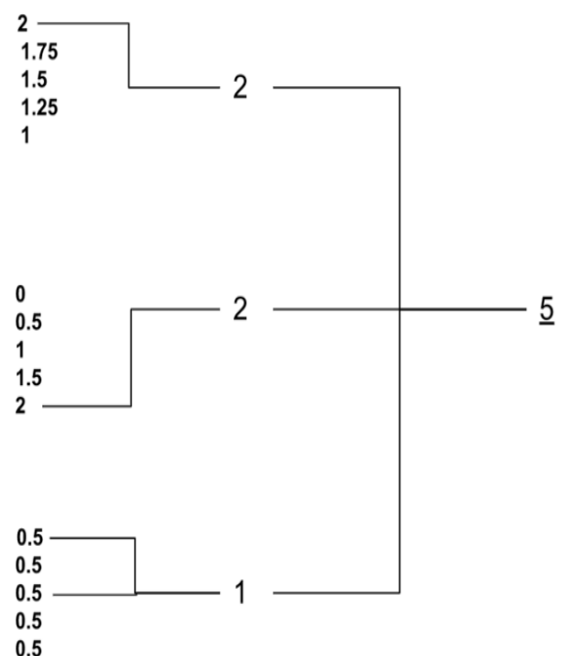


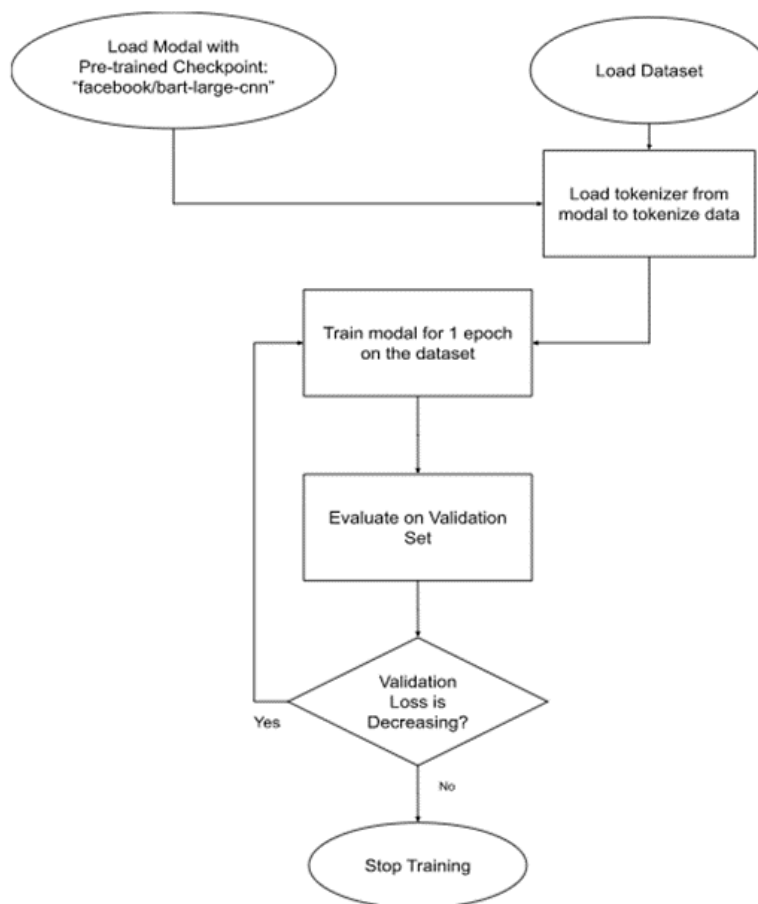
Figure 6: Example calculation of severity score based on answer selected.

The positive sentiment is considered as 0 and higher scores indicates severe symptoms. According to the options selected by the patients, the total score will be calculated, and the severity is given Red (greater than 20), Yellow (between 10 and 20) and Green (less than 10).

## Fine-tuning BART for questionnaire summarization

We have used [BART](#) model to provide the summarization of any questionnaire that would be used in order to filter the patients based on the therapist who is using it. They will receive an automated email with summary of the respondent(Lewis et al., 2019).

The model is used is a pre-trained Model - [BART](#) checkpoint on CNN-DM dataset. It was trained again in SAMSum dataset with early stopping so that it could provide summary for dialog. The evaluation metrics used include loss and ROUGE scores for both validation and test set.



*Figure 7: Finetuning process of BART on SAMSum dataset*

The above flowchart gives an overview on how we trained the model so that it can give us the summarization of the questionnaire response. At first the pre-trained model with the checkpoint: "facebook/bart-large-cnn" is loaded along with the dataset SAMSum. They are both fed into the tokenizer model to tokenize the data. After that the model is trained with 1 epoch on the dataset, the validation set is evaluated. If the validation is decreasing, we again

train the model with 1 epoch but if the validation loss has stopped decreasing, we stop the training.

We have used the following evaluation metrics for validation set:

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

**ROUGE, or Recall-Oriented Understudy for Gisting Evaluation (Lin, 2004)**

**ROUGE-N** measures the number of matching n-grams between the model-generated text and a human-produced reference. An n-gram is simply a grouping of tokens/words. A unigram (1-gram) would consist of a single word. A bigram (2-gram) consists of two consecutive words.

**ROUGE-L** is based on the longest common subsequence (LCS) between our model output and reference, i.e., the longest sequence of words (not necessarily consecutive, but still in order) that is shared between both. A longer shared sequence should indicate more similarity between the two sequences.

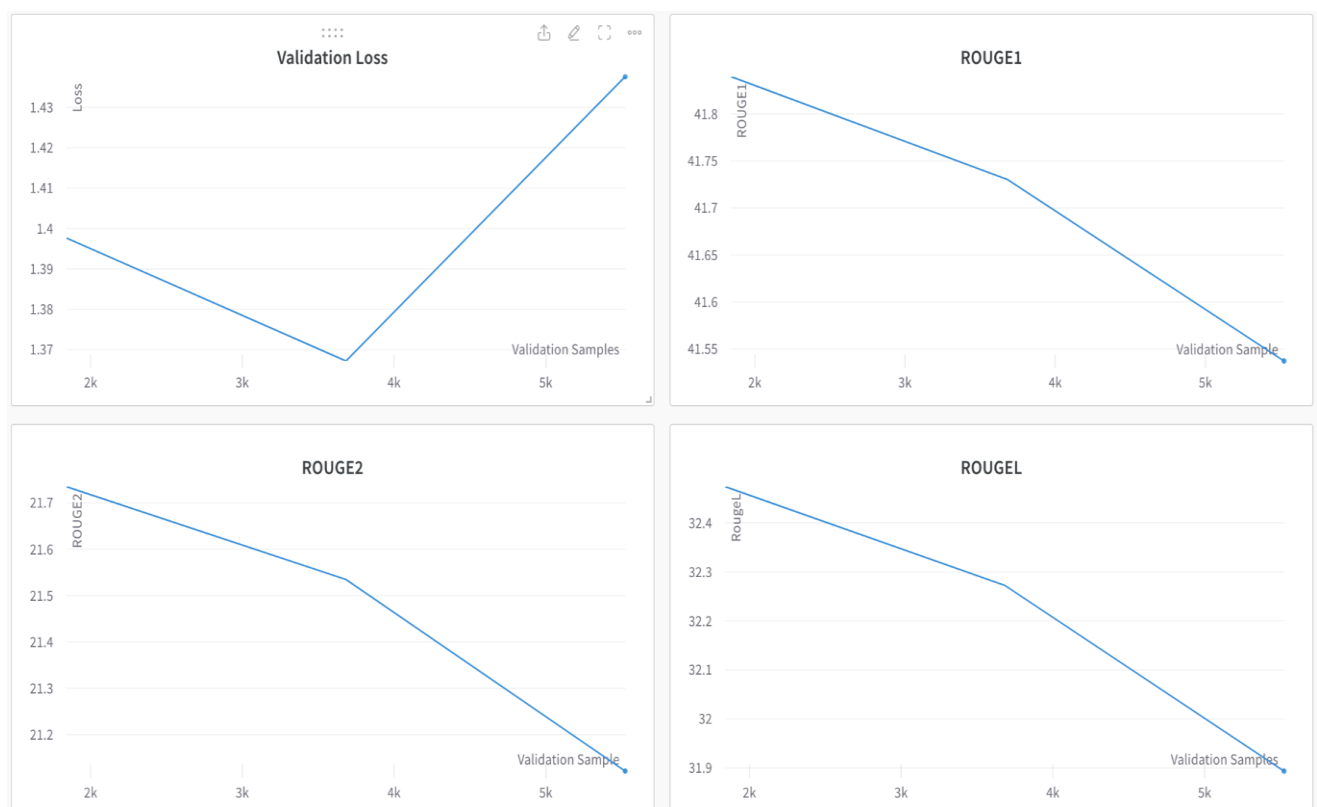


Figure 8: Evaluation metrics at each epoch

## Evaluation of Validation Set

*Table 3: 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 4:: 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:

The mathematical scoring logic model based on the questionnaire that we created in our initial phase that gave us a severity report after any participant fills up the form. The report consists of the demographics of the patient. There is gender icon that indicates the gender of the patient. There is a history section that talks about any previous diagnosis, medication, or treatments. The assessment section indicates the current need of the patient.



### Patient Summary

● Severity: Orange

### Demographics



Initials: rf  
Gender: Woman  
Age: < 18  
Country-Province/State: Canada, British Columbia  
Occupation: Student

### History

Had previous diagnosis: No  
Experienced symptoms: N/A  
Treatment Received in past: N/A  
Receiving any treatments: N/A  
Medication history: N/A

### Assessment Needs

Condition: Substance Abuse  
Affect on daily life: N/A  
Affect on sleep pattern: Moderately  
Last experienced symptoms: < 6 months  
Coping Mechanism: cycling  
Preferred Treatment:  
Comfortable Treatment: Yes


### About CogniXR

CogniXR is an easy-to-use HIPAA-compliant platform that mission is to expand the reach of mental health care so that wellness programs and health care providers support more people on their recovery journeys without delay.

[Ask us a question](#)

Figure 9: Summary report with severity

The second results of the BART model is the summary of the questionnaire that is fed into the model. It creates a summary of the overall content of the response and has both English and French section. It is as shown in the following image.



**cogniXR**  
HEALTH

### Work Stress Questionnaire Response Summary

- Tina works as Business Data Analysis
- She feels overwhelmed by her work and stressed about her work
- Tina's organization has a mental health counselor
- Tina would like to seek help from mental health expert if she feels she cannot balance work and personal life
- Tina is not comfortable with talking about her well-being with her team or manager

### Work Stress Questionnaire Résumé de la réponse

- Tina travaille dans l'analyse des données d'entreprise
- Elle se sent dépassée par son travail et stressée par son travail
- L'organisation de Tina a un conseiller en santé mentale
- Tina aimerait demander l'aide d'un expert en santé mentale si elle sent qu'elle ne peut pas concilier travail et vie personnelle
- Tina n'est pas à l'aise de parler de son bien-être avec son équipe ou son manager

Note: This content is generated using BART, a Large Language Model. This content is not moderated and might be offensive, repetitive or might contain additional information generated by AI.

### About CogniXR

CogniXR is an easy-to-use HIPAA-compliant platform that mission is to expand the reach of mental health care so that wellness programs and health care providers support more people on their recovery journeys without delay.

[Ask us a question](#)

Figure 10: Summary report with BART

# Conclusions and Recommendations

During our project we made some assumptions at the beginning, which were the basis of our project. As we used scoring logic to address the answers in such a way that it can determine the scale of each answer's weight. The first assumption we made was that the mathematical scoring logic is correct based on the. The second assumption was that as the scores are added to a total, they will give us the severity score that is based on Severity determined by total score added. The third assumption was that the questions and answer are in the form of conversation between two people like an interrogation. Also, our client wanted a translation of English to French as in CogniXR is based on Montreal, both languages are used. For the completion of this requirement, we assumed that translation does not cause data loss.

However, we have to consider the biases in the project. One of them is that missing out on useful data as we are using summarization method, the model is not sensible enough for the sensitive content. The second one is there is no filtration of the offensive content as there are chances of participants to fill offensive contents which might interfere with the summary. The last one is the chances of creating duplicate sentences to fulfil minimum word criteria. The criterion of our model is set for 10 to 30 words summary for every 100 words. But as we are unsure of the number of words for summary generation just to fulfil the criteria duplicate sentences might be created as a result.

Likewise, we must address the challenges in our project are summarization bias as there can be compromised by duplication of the sentences. Also, there can be inaccurate categorization of the patients as there is no filtration of the offensive contents which can interfere in categorization. Also scoring model may give unintended output as the weight we have assigned to each answer for each question may not reflect the actual weight of the impact on the patient.

For the further improvement on the product, the severity level for every type of mental health patient intake questionnaire based on every therapist. This can be done by using word-based model from the response. For example, if the response includes any word indicating self-harm or suicide or usage of medication the severity will be red automatically. This technique can be similar to the sentiment analysis which is used to summarize the overall sentiment of the content by usage of Natural Language Processing, text analysis and so on.

We have to consider the biases in the project. One of them is that missing out on useful data as we are using summarization method, the model is not sensible enough for the sensitive content. The second one is there is no filtration of the offensive content as there are chances of participants to fill offensive contents which might interfere with the summary. The last one is the chances of creating duplicate sentences to fulfil minimum word criteria for the summary.

Our first approach can provide severity of patients for mental health condition and second approach can summarize questionnaire filled by the patients saving providers' time. In future, methods can be developed to combine both approach which would give summary as well as severity for any psychometric questionnaire. This would require preparation of dataset where severity is manually annotated for a response. An NLP model would then be able to identify key phrases and features determining severity and classify the response accordingly.



Further on if the severity is either orange or red the patients can be redirected to second phase of patient analysis. This can be done by matching each condition with the therapy that they require which can then go to the emotion detection step which will help the therapist to provide the required form of therapy and analyse if their session made any impact on the patient.

## Acknowledgment:

We would like to acknowledge and express our gratitude to our supervisor. Mr. Umair Durrani, Professor of St. Clair College, for his help, stimulating suggestions and encouragement throughout the execution of the Capstone project.

We would like to thank Ms. Emmanuella Michel, Founder of CogniXR Health, for providing us with this opportunity to collaborate with the organization. She guided us throughout the project such that we could fulfil her requirements as instructed.

We are also grateful to St. Clair College for giving us this exposure to work in a real-life environment which has helped us to gain knowledge and experience. This has helped us to polish our skills to some extent that can help us grow more in the future.

We also extend our warmest regards and gratitude to all the participants and our classmates who took part in our survey data collection and supporters who have helped us in any way throughout the course of the project.

## References:

1. *10 facts on mental health*. (n.d.). Retrieved April 24, 2023, from <https://www.who.int/news-room/facts-in-pictures/detail/mental-health>
2. *Center for Epidemiological Studies Depression (CESD)*. (n.d.). <https://www.apa.org>. Retrieved April 20, 2023, from <https://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/depression-scale>
3. *HAMILTON-DEPRESSION.pdf*. (n.d.). Retrieved March 18, 2023, from <https://dcf.psychiatry.ufl.edu/files/2011/05/HAMILTON-DEPRESSION.pdf>
4. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension* (arXiv:1910.13461). arXiv. <https://doi.org/10.48550/arXiv.1910.13461>
5. *Mental Health—PAHO/WHO | Pan American Health Organization*. (n.d.). Retrieved March 28, 2023, from <https://www.paho.org/en/topics/mental-health>
6. *Penn State Worry Questionnaire (PSWQ)—Nathan Kline Institute—Rockland Sample documentation*. (n.d.). Retrieved April 28, 2023, from [http://fcon\\_1000.projects.nitrc.org/indi/enhanced/assessments/PSWQ.html](http://fcon_1000.projects.nitrc.org/indi/enhanced/assessments/PSWQ.html)
7. *PTSD Checklist for DSM-5 (PCL-5)—PTSD: National Center for PTSD*. (n.d.). Retrieved Feb 28, 2023, from <https://www.ptsd.va.gov/professional/assessment/adult-sr/ptsd-checklist.asp>
8. *Samsum · Datasets at Hugging Face*. (2022, December 23). <https://huggingface.co/datasets/samsum>
9. Saxena, S., Thornicroft, G., Knapp, M., & Whiteford, H. (2007). Resources for mental health: Scarcity, inequity, and inefficiency. *The Lancet*, 370(9590), 878–889. [https://doi.org/10.1016/S0140-6736\(07\)61239-2](https://doi.org/10.1016/S0140-6736(07)61239-2)
10. Weathers, F. W., Marx, B. P., Friedman, M. J., & Schnurr, P. P. (2014). Posttraumatic Stress Disorder in DSM-5: New Criteria, New Measures, and Implications for Assessment. *Psychological Injury and Law*, 7(2), 93–107. <https://doi.org/10.1007/s12207-014-9191-1>
11. **Link to the copy of Questionnaire used for the project:**  
[https://forms.office.com/Pages/ShareFormPage.aspx?id=b2eGyTmbCE20-KZo4OjGpai-ewZ\\_rJZAnsCyF4k7Ti1UMjJGOU1UT1pTNEJaMjg1VzQ5T09RVTdCUy4u&sharetoken=pdRXMss97gtAB4su4JKB](https://forms.office.com/Pages/ShareFormPage.aspx?id=b2eGyTmbCE20-KZo4OjGpai-ewZ_rJZAnsCyF4k7Ti1UMjJGOU1UT1pTNEJaMjg1VzQ5T09RVTdCUy4u&sharetoken=pdRXMss97gtAB4su4JKB)
12. **Link to Github:** [https://github.com/AabritiKarki/CapstoneProject\\_Group8](https://github.com/AabritiKarki/CapstoneProject_Group8)