ECE 1786 Final Report

Group Member:
Bill Zou
student number:1004761778
Email: zouyuang@mail.utoronto.ca
Shiyu Xiu
student number: 1004724872
Email: shiyu.xiu@mail.utoronto.ca

Word Count: 1993

Introduction

The goal of this project is to apply NLP techniques to analyze text responses to an open-ended survey question in order to provide a quick overview of all the responses received. Specifically, we focus on analyzing responses received from a safety training survey conducted by CAMH. Due to the large number of responses received, we believe our project can greatly improve the efficiency of reviewing those responses by providing meaningful insights for the inspector.

Illustration/figure

Our implementation contains two fine-tuned deep learning models: a BART-based summarizer and a BERT-based classifier. The figure below shows the overall structure of our project.

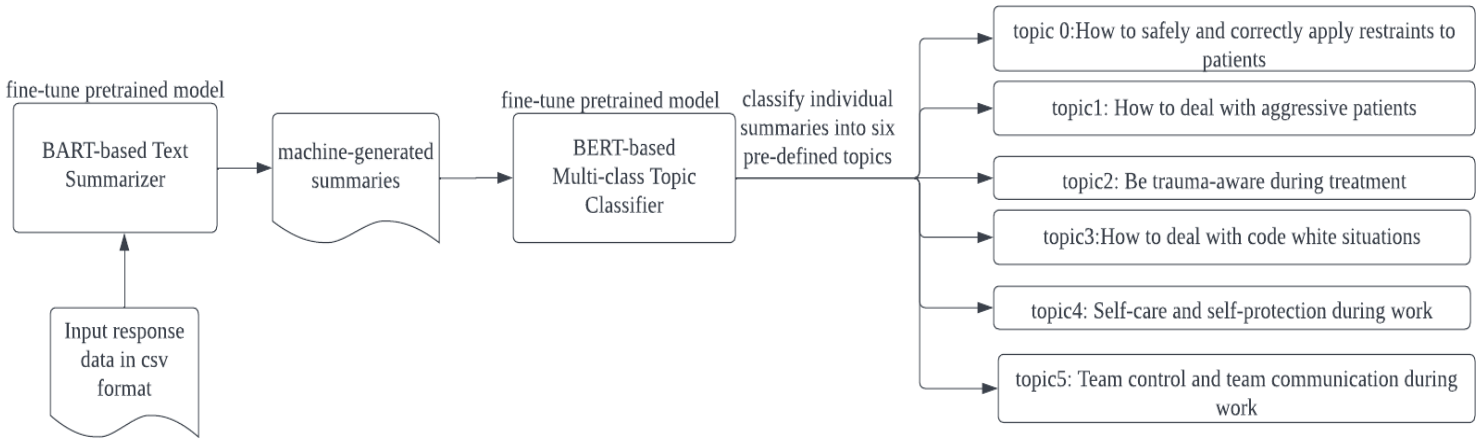

Figure1: General Structure of our Project: Input responses are summarized and then classified into pre-defined topics.

Background & Related Work

This section describes related work and software in the field of Topic Modelling as well as Text Summarization.

- **BertTopic(Topic modelling):**
    - BertTopic (Grootendorst,2022) is a clustering technique to explore latent topics in a collection of documents. BertTopic models the topics by extracting the topic representation using a variation of TF-IDF. Specifically, the process involved three steps. It first generates document embedding using pre-trained transformer-based language models. Secondly, It performs dimensionality reduction to improve the clustering process later on. Lastly, the topic representations are chosen from the document sets using a class-based variation of TD-IDF.
    - We decided to use BertTopic as an assisting tool for this project. Some of the topics generated by BertTopic help us to come up with the six pre-predefined topics for the classifier.
- **GPT-3 from OpenAI:**
    - OpenAI GPT-3 summarizer is trained by summarizing books. The model first summarizes small sections of a book and then further summarizes those summaries into higher-level summaries. GPT-3 is able to provide higher quality summaries than other large models such as T5 and BART since reinforcement learning from human feedback is included when training the GPT3 model. For each summarization generated, a human-made label is assigned in order to build a classifier that predicts whether the summary is aligned with human preferences. Large summaries are decomposed into several shorter pieces for humans to label them easily. By training the classifier, GPT3 is able to produce summaries that are highly similar to human-generated ones.
    - We have tried zero-shot in GPT3 with some single responses concatenated to imitate a large response. However, the generated summary compared with the input response, leads to severe information loss.

Data and Data Processing

The dataset we use for this project is provided for us by the Center of Addiction and Mental Health(CAMH). This dataset includes 550 text responses to an open-ended question “How do you intend to apply the safety training you received?” An example of the original dataset before processing is shown below.

Intent2Use_Open
.
Self care during work and before and after (deep breathing, positive motivation, olfactory). Sa
.
.
.
In the applications on mechanical restraints and in management of aggressive behaviors
Apply same type of restraint to wrists/hands as to foot. Adjust tightness so not a chooking hazard
Properly putting restraints on i.e. feet together closer
1) Completing more casual client debriefs as oppose to just seeing it as a task to check off 2) More motivated to implement self-care strategies
Practice
Helping client to complete a debrief session
.
Reminder of trauma-informed care. It was nice to have the service user educator present for a different perspective

Figure 2: Original dataset before processing

After receiving this dataset, we perform data processing as follows:

1. Filter out empty responses.
2. Perform decontraction (we'll -> we will, I can't -> I cannot etc.)
3. Perform train/evaluation/test split.

An example of the dataset after processing is

	Intent2Use_Open
0	Self care during work and before and after (deep breathing, positive motivation, olfactory). Sa
1	In the applications on mechanical restraints and in management of aggressive behaviors
2	Apply same type of restraint to wrists/hands as to foot. Adjust tightness so not a chooking hazard
3	Properly putting restraints on i.e. feet together closer
4	1) Completing more casual client debriefs as oppose to just seeing it as a task to check off 2) More motivated to implement self-care strategies
5	Practice
6	Helping client to complete a debrief session
7	Reminder of trauma-informed care. It was nice to have the service user educator present for a different perspective
8	1. Applying restraints and removing in way instructed today. 2. Trying to manage the client to try and prevent the use of mechanical restraints.
9	Participate with more confidence in code whites and proper hand-hold technique
10	Safer way of using restraint and self-care
11	shoulder strap on head rest of bed
12	1) Debrief meetings right after restraint event/Code white 2) Self-care - create a plan (long/short term)
13	Retraining events at work; setting up restraint work to improve flow at work

Figure3: Original dataset after processing

We run the pre-processed dataset through BertTopic and obtained the distribution of topics across the responses as follows:

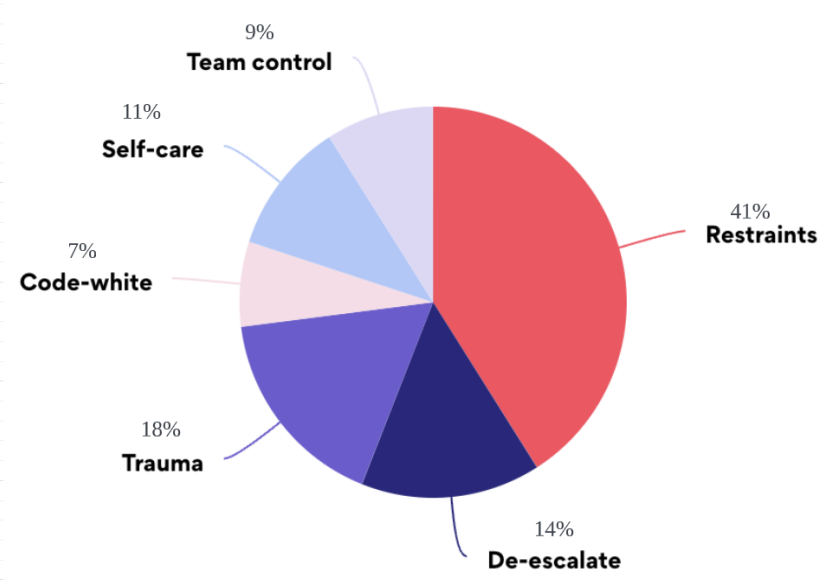

Figure4: Distribution of Topics in dataset

We notice that the dataset is imbalanced, with a majority of people responding under the topic “Restraints”. Also, we observe that some of the responses are irrelevant such as: ‘thank you’, ‘this is a good training’ or ‘I will apply it’. We decide to keep these responses as they reflect what read-world data would be like.

Architecture and Software

There are two large pre-trained models from huggingface in our project. One for summarizer and the other one for the classifier.
The model used for the summarization task is the facebook/bart-large-xsum, the hyperparameters used in the trainer are learning_rate=2e-5,batch_size = 5, Num_Epochs = 10 and the total number of parameters is 406 million.
For the classification task, we used distilbert-base-uncased. The hyperparameters used in the trainer are learning_rate= 5e-5, batch_size= 8, Num_Epochs = 6 and the total number of parameters is 109 million. To reproduce a model similar to ours, you can load these models from huggingface and finetune them with your own dataset. Please mind that you may need to finetune the hyperparameters as well.

Baseline Model

- **TextRank:** TextRank is our baseline model for summarization. TextRank is an extractive and unsupervised method inspired by the PageRank algorithm. Since TextRank is a non-machine learning approach, we didn’t do any training for it.
- **Decision Tree:** The decision tree from sklearn is the baseline model for classification. A Decision Tree is a tree-structured supervised model utilized for classification and regression tasks. We trained a decision tree with depth =

5 using 101 responses. These responses are classified into 6 pre-defined labels.

Quantitative Results:

For the topic classification task, we utilized the average accuracy for both the baseline model and the main model. For the decision tree baseline model, we achieved an average accuracy of 54%. While for our main transformer-based model, we achieve an average accuracy of 68%.
For the summarization task, we utilized the rogue score as a proxy metric for comparison. We have a baseline with Rogue-1: 0.85, Rouge-2: 0.83, Rogue-L: 0.75, while for the main model, we have Rogue-1: 0.45, Rouge-2: 0.18, Rogue-L: 0.31. We obtained a higher rogue for baseline(extractive algorithm) since rogue is comparing the overlapping words between the input and reference. Our main model is doing abstractive summary, so the overlapping should be lower and therefore, score is lower.
Moreover, we used the BertScore to further evaluate our main models and achieve a pretty decent score using BertScore. We obtained average precision of 0.8926262974739074, an average recall of 0.8879761099815369 and an average f1 of 0.8902641654014587 for our summarizer.

Qualitative Results

In this section, we include some sample inputs/outputs that illustrate when our model is working well as well as when it is not.
Summarizer:

- **Input responses:** *a: Be more considerate of client having PTSD some kind of trauma. Go back to appropriate self-care activities geared to me Continue to make every effort to de-esclate using verbal skills. Always operate from a trauma-informed practice Using the "more trauma informed" skill before deciding on use of seclusion room. Continue to provide debrief assessment to client even after they refuse. Trauma informed care, de-escalation methods/approach, and least/last option/restraints Trauma strategey I will work harder to debrief with team and clients after an incident. I will incorporate trauma awareness to my decision making in regards to applying restraints as a last resort. Trauma informed care like talking clients through things, not staring, covering them Trauma-informed care and de-escalation To assess risk of harm, engage the client and plan team intervention in violent emergent situations Trauma informed, self-care*
- **Baseline result:** *I will incorporate trauma awareness to my decision making in regards to applying restraints as a last resort.Trauma informed care, de-escalation methods/approach, and least/last option/restraints Trauma*

strategey I will work harder to debrief with team and clients after an incident.Trauma informed care like talking clients through things, not staring, covering them Trauma-informed care and de-escalation To assess risk of harm, engage the client and plan team intervention in violent emergent situations Trauma informed, self-care. Continue to provide debrief assessment to client even after they refuse. Go back to appropriate self-care activities geared to me Continue to make every effort to de-esclate using verbal skills.

- **Main model result:** *Be more considerate of client having some kind of trauma. Always operate from a trauma-informed practice. De-esclate using verbal skills. Use the "more trauma informed" skill before deciding on use of seclusion room. Work harder to debrief with team and clients after an incident*
- **Comparison:** Our main model is able to generate a more concise summary compared with the baseline. Also, the baseline result contains a lot of long and hard-to-read sentences while the sentences in main model result are shorter and more effective.

Classifier:

- **Example Input Response:** *“Using more hand controls and scanning all beds for landmarks and applying extender to limb for better restraint application.”*
- **Baseline result:** *label 1 -> “How to deal with aggressive patients”*
- **Main Model result:** *label 0 -> “How to safely and correctly apply restraints to patients”*
- **Comparison:** The main model is able to predict the correct label while our baseline offers an incorrect prediction.

However, there are also cases when our main model is not performing well. Specifically, we notice that summarization sometimes causes information loss compared with the original responses. Also, the classifier can predict a wrong label for a response. For example:

Summarizer:

- **Input responses:** *a: Teaching new staff (providing support). Be more considerate of client having PTSD some kind of trauma. Go back to appropriate self-care activities geared to me. Invovled more in code white situation. Trauma-informed care and de-escalation.Training gives me good understanding when to apply restraint Level 1 physical restraint when transferring patient and hand control.*
- **Main model result:** *Teaching new staff support. Be more considerate of client having trauma. Invovled more in code white situation. Go back to appropriate self-care activities. Train more in trauma-informed care and de-escalation.*
- **Comment:** We see that generated summary contains loss in information. The last sentence in input is ignored by our model and not included in the summary.

Classifier:

- **Example Input Response:** *“Protect agitated clients and staff from harm.”*
- **Main Model result:** *label 1 -> “How to deal with aggressive patients”*
- **Correct result:** *label 4 -> “Self-care and self-protection during work.”*
- **Comment:** We see that for this response, our classifier is not able to offer a correct prediction. It classifies the response into “How to deal with aggressive patients” since it sees “agitated”. However, this response focuses on protection during work.

Discussion and Learnings

Figure 5: The final output of SurveyInsider

Index	How to safely and correctly apply restraints to patients ▲	How to deal with aggressive patients	Be trauma-aware during treatment	How to deal with code white situations	Self-care and self-protection during work	Team control and team communication during work
Percentage	22%	25%	25%	9%	3%	16%
summary	Use PINEL discontinuation strategies more effectively / More hands on practice with restraints / The training gives me more confidence and knowledge how to do mechanical restraint Pin at the middle during restraint event / Use seclusion more than restraints / Better understanding of when to apply restraint / Refreshed learning of limb restraint / One person only take to patient when applying restraints.	Be mindful of effect of our actions has on clients / Manage agitated patient / Debrief after restraint / Protect the client from harm when they are agitated / Protect other client and staff of possible harm from an agitated client / Applying restraints and how to hold aggressive client / When client needs to be controlled / To client de-escalation	De-escalate trauma-informed care and team communication / Remember trauma-informed practice and apply restraints in the correct way / Use self-care and be more aware of trauma informed care / Be mindful trauma can make restraints event worst for patient / Trauma-informed de-esclation prior to use of restraints. / Be considerate of client having PTSD some kind of trauma / Use trauma-informed practice before	More use of Red Zone and more in code white situation. / Do code white and de-escalation / Every code situation	Go back to appropriate self-care activities geared to me	Use technique to work with team and discuss strategies learnt / Encourage team on self-care and support team on restraints / Create a plan / Use training when client becomes indirectable. / Make every effort to de-esclate using verbal skills

Our results seem quite reasonable on the test data as shown in figure 5(the column is the pre-defined topics; rows are percentages as well as the summaries under each topic). We learned the power of fine-tuning the transformer-based pre-trained model. Even with just a small set of training data, the model is able to successfully perform the tasks.

For improvement, firstly, we can improve the way of training the summarizer. We included some irrelevant reponses in its training data since we want it to learn how to ignore those responses when generating the summary. However, this may cause the summarizer to ignore relevant responses as well. An alternative way is to use a classifier to filter out irrelevant responses, then use the remaining responses as the training data for summarizer. Secondly, we could do supervised topic modelling instead of classification. Using supervised topic modelling enables us to extract topics from the responses automatically. In this case, if we have a response that doesn’t belong to any of the pre-defined topics, we can still assign a correct topic to it.

Individual Contributions

Throughout the semester, we think we have formed a really effective and cooperative team. Since we have been working together in person and Github does not support collaborating on notebook files, therefore, the commits might not reflect what we did indeed.

Shiyu:

- Did the data preprocessing: Decontraction
- Research on Topic Modelling
- Performed topic modelling on cleaned dataset
- Labeled half of the data
- Built and evaluated the baseline for the summarization task
- Finetuned main models: both summarization and classification task with different hyperparameters
- Researched on rouge score and implemented rouge measurement for summarization task
- Designed the table structure to effectively show both summarization and classification results to user
- Implemented the Gradio UI

Bill:

- Did the data preprocessing: Remove empty responses
- Research on Text Summarization
- Perform analysis on the results of topic modelling and splitted dataset by topics for labelling
- Labelled the other half of the data
- Performed train/eval/test splits
- Built and trained the baseline for the classification task
- Built the code structure for finetuning main models: both summarization and classification task
- Researched on bert score and implemented bert measurement for summarization task
- Automated the process of generating the final output.

Reference:

[1] M. Grootendorst, “Bertopic: Neural topic modelling with a class-based TF-IDF procedure,” [Accessed: 11-Mar-2022].

[2] “Text summarization,” *Amazon Science*. [Online]. Available: https://www.amazon.science/tag/text-summarization. [Accessed: 13-Dec-2022].

[3] K. Wiggers, “OpenAI unveils model that can summarize books of any length,” *VentureBeat*, 23-Sep-2021. [Online]. Available: https://venturebeat.com/business/openai-unveils-model-that-can-summarize-books-of-any-length/#:~:text=A%20fine%2Dtuned%20version%20of,calls%20%E2%80%9Crecursive%20task%20decomposition.%E2%80%9D. [Accessed: 13-Dec-2022].

Permissions

Permission to share:	Video	Report	Source Code
Bill Zou	Yes	Yes	No
Shiyu Xiu	Yes	Yes	No