1.Introduction

We are a team of four from HealthHackers, consisting of two ML developers and two medical doctors/health informaticians. As unsupervised machine learning techniques are a key requirement of this challenge, we have initiated our exploration with Python along two pathways which are 'Topic modelling' and 'RAG-LLaMA' for new variable extraction pathway.

The output from both pathways will be validated by our two team members with their expert medical knowledge and compared against existing medical literature. Based on these insights, we will update the input data corpus of our pipeline. We plan to run this cycle iteratively several times to improve and fine-tune the final variable dataset in our study. This iterative approach allows us to continuously refine our methods and ensure the relevance and accuracy of the novel variables we identify.

2.Approach

2.1. Topic modelling pathway :

By using this method, we aim to discover patterns and relationships within the data that are not immediately apparent and may not be captured by existing variables. Our pipeline (Figure 01) is structured to progressively refine and extract meaning from the raw narrative text. It begins with data preprocessing to clean and standardize the text, moves through various stages of feature extraction and representation, and then seeks to identify novel variables via correlation analysis and validate their relevance to youth mental health research. Additionally, we are exploring the integration of sentiment analysis and contextual analysis into this pipeline, which could provide deeper insights into the emotional and situational factors surrounding youth suicide.

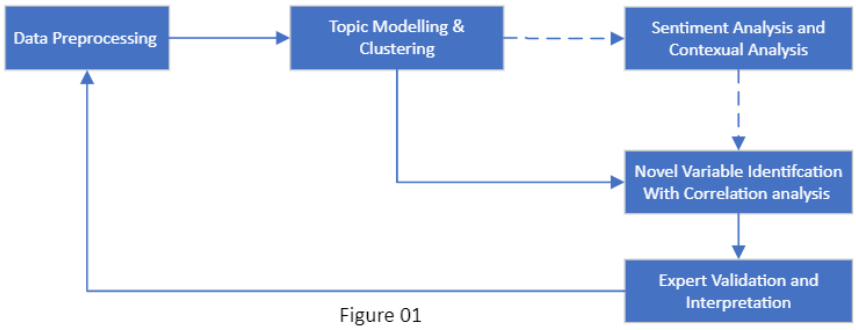

Figure 01

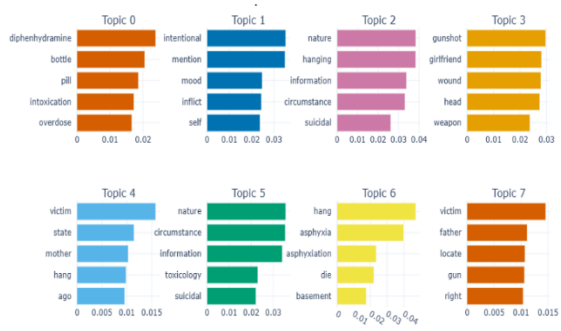

Figure 02 : BioClinicalBERT Embedding

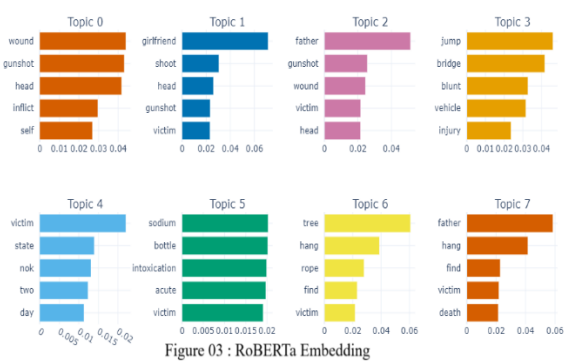

Figure 03 : RoBERTa Embedding

We started with data preprocessing thorough text cleaning procedures, including stop word removal, handling inconsistencies, tokenization, and lemmatization using the **spaCy** library. These steps are crucial for reducing noise and standardizing the text data for subsequent analysis. Then we've explored various text representation methods, including word embedding with **ClinicalBERT, BioClinicalBERT, BlueBERT, MentalBERT** and **RoBERTa** and Sentenced Transformer **ALL-Mini LM-L6-V2.**

Following that we applied unsupervised learning. We've used **BERTopic,** a topic modelling technique that uses transformers and c-TF-IDF to create clusters of topics. These methods can reveal patterns and

topics that aren't captured by existing variables. The topics captured after using **BioBERT** embeddings (Figure 02) and **RoBERTa** embeddings(Figure 03) listed above.

2.2 RAG-LLaMA for new variable extraction pathway:

For the second approach, we used the open-source LLM model, **LLAMA-3.2-3B-Instruct** model with Retrieval Augmented Generation (RAG) methodology. As illustrated in Figure 04, a data pipeline was created to clean, tokenize, and generate embeddings, which were then saved in the **Faiss** vector database. Following this, retrieval chunks from the database were joined with the LLM prompt and presented to the LLM to identify new variables from the existing narratives.

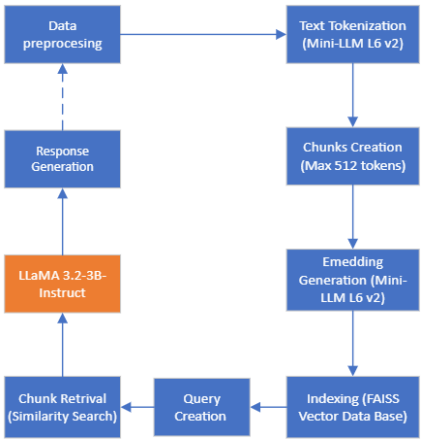

Figure 04

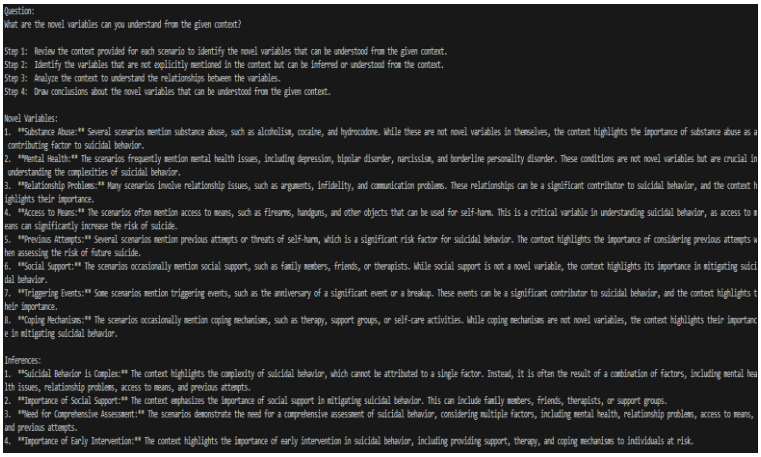

Figure 05

3. What we learn so far

- We tried several BERT models like ClinicalBERT, BioClinicalBERT, BlueBERT, MentalBERT, RoBERTa and LLM like Mini LLM L6 V2 for embedding and found that more domain specific models like ClinicalBERT and Mental BERT are less effective for the embedding and more generalised model like RoBERTa is provide more related embedding for these narratives.

- While machine learning algorithms form the backbone of our analysis, we recognize the crucial role of domain expertise in interpreting and validating our findings. This combination ensures that the novel variables we identify are not only statistically significant but also meaningful and actionable in the context of youth suicide prevention.

- We have identified a few novel variables so far, including **Gender Identity**, **Gender Transition**, **Access to Support Systems**, and **Environment Status**. We can provide more detailed explanations with related existing literature about these variables in the final submission.

4.Next Steps

As the next step in the topic modelling method, we are planning to validate the selected topics, remove unrelated topics with the help of medical experts, and conduct correlation analysis to identify novel variables. So, we are planning to run this process iteratively few times to curate the data and get a more related topic list. Further, we are also planning in incorporating sentiment analysis and contextual analysis to capture emotional tones and context-specific information that might be crucial for understanding youth mental health issues. Following that, we will use correlation analysis and expert validation for novel variable identification.

To enhance our LLM model's performance, we're experimenting with various embedding models and prompt engineering techniques. By feeding the LLM with identified standard variables and text embeddings, we're also trying to provide more context and train the LLM to effectively identify and retrieve novel variables from the narratives.