Introduction

We are a team of four from HealthHackers, consisting of two ML developers and two medical doctors/health informaticians. We applied advanced unsupervised machine learning techniques using **Python's program** to systematically address the challenge of extracting novel variables from narrative suicide data.

Methodology

Our methodology for identifying novel variables related to suicide from the National Violent Death Reporting System (NVDRS) narrative text data is founded on two fundamental principles: domain expertise integration and scalable architecture design.

The first principle emphasizes the critical role of human expertise in healthcare analytics. We incorporate domain expert knowledge throughout our pipeline with the involvement of two medical doctors in our team, from initial data preprocessing to final validation of results. This human-in-the-loop approach is particularly crucial in healthcare applications, where misinterpretation or oversights could have significant consequences. Expert involvement serves multiple functions: validation of extracted variables, identification of potential biases or inconsistencies, and ensuring clinical relevance of the findings. This approach aligns with recent research suggesting that hybrid human-AI systems often outperform purely automated solutions in healthcare settings(1). The second principle focuses on developing a scalable and adaptable framework that extends beyond suicide-related variable extraction to other healthcare domains. This emphasis on scalability ensures that our methodology can be effectively applied to various healthcare analytics challenges, from mental health to broader public health surveillance.

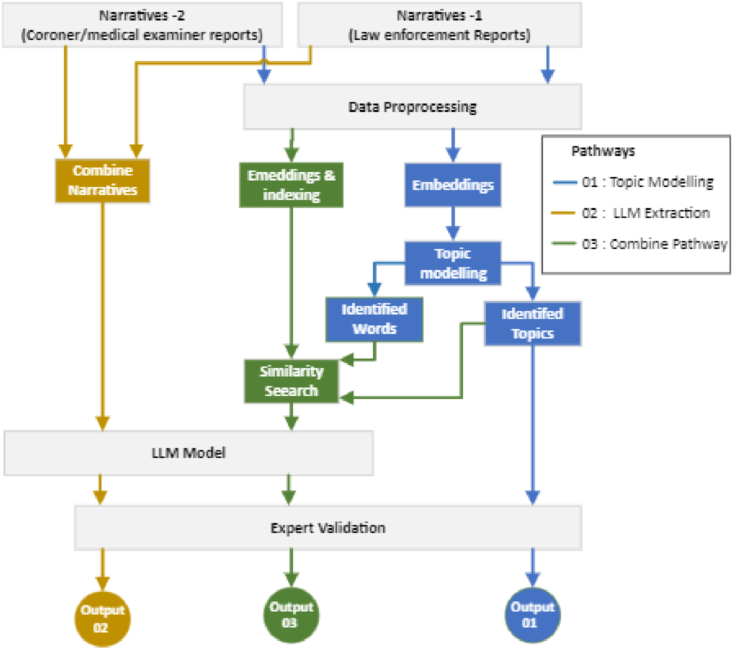


Figure 01: Data Pipeline (High-Level Workflow)

As illustrated in Figure 01, we developed a Python-based pipeline-with three interconnected pathways for extracting variables from the NVDRS dataset: a 'Topic Modeling pathway' for identifying latent themes, a 'Large Language Model (LLM) Responses Clustering pathway' for direct variable extraction, and a 'Combined pathway' that synthesizes outputs from both approaches. Each pathway's output underwent rigorous validation by two medical doctors who independently evaluated and reached a consensus on clinically relevant variables. This dual-review methodology, followed by consensus determination, ensures both the clinical validity of the selected variables and minimizes individual bias in the selection process.

Pathway 01: Topic Modeling

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 (Figure 02). Text preprocessing employed **spaCy** including tokenization, removing stop words, lemmatization, and handling linguistic inconsistencies to standardize the narrative text for subsequent analysis (Code Block 01). The cleaned text was normalized into a string for embedding and topic modeling. We evaluated several domain-specific transformer models including ClinicalBERT, BioBERT, and LegalBERT for text embeddings, but found that **RoBERTa** and **All-MiniLM-L6-v2** outperformed these specialized models due to the predominantly general-language nature of the suicide narratives, which contained limited medical and legal terminology. Therefore, **RoBERTa** and **All-MiniLM-L6-v2** were selected for embeddings. **BERTopic** analysis generated 52 distinct topics and their associated word clusters from the narrative data. Following the medical expert's validation, clinically relevant topics were transformed into novel variables, while selected word clusters were integrated into our combined pathway.

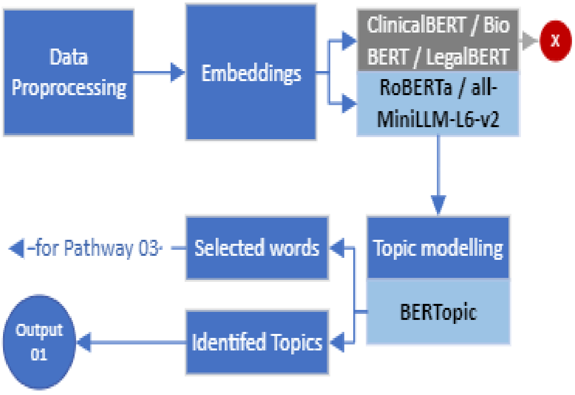
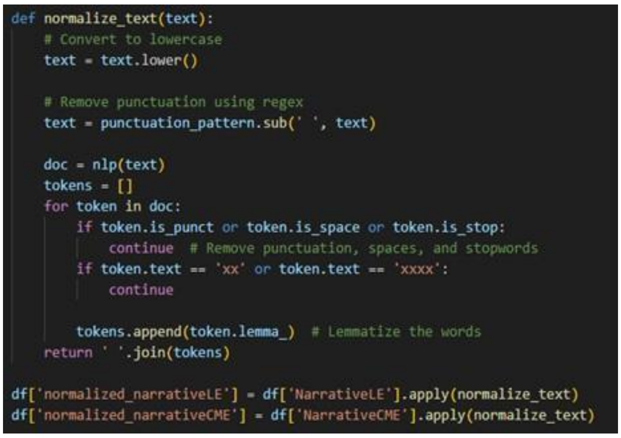


Figure 02: Topic Modeling Pathway



```python
def normalize_text(text):
    # Convert to lowercase
    text = text.lower()

    # Remove punctuation using regex
    text = punctuation_pattern.sub(' ', text)

    doc = nlp(text)
    tokens = []
    for token in doc:
        if token.is_punct or token.is_space or token.is_stop:
            continue  # Remove punctuation, spaces, and stopwords
        if token.text == 'xx' or token.text == 'xxxx':
            continue

        tokens.append(token.lemma_)  # Lemmatize the words
    return ' '.join(tokens)

df['normalized_narrativeLE'] = df['NarrativeLE'].apply(normalize_text)
df['normalized_narrativeCME'] = df['NarrativeCME'].apply(normalize_text)
```

Code Block 01

Pathway 02: LLM Responses Clustering

For our second pathway, we leveraged the capabilities of Large Language Models, specifically implementing LLAMA-3.2-3B-instruct, an open-source model chosen for its robust performance in instruction-following tasks and text analysis. Our objective was to extract suicide-related variables from narrative texts using LLM by implementing carefully engineered prompts (Figure03).

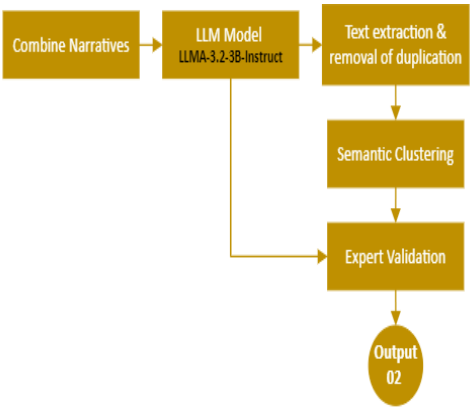
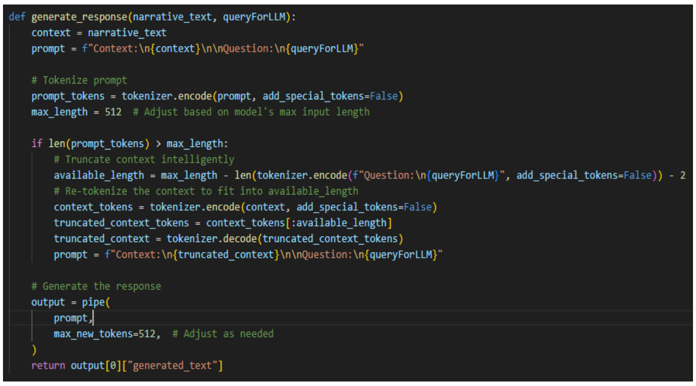

Combine Narratives
LLM Model LLMA-3.2-3B-Instruct
Text extraction & removal of duplication
Semantic Clustering
Expert Validation
Output 02

Figure 03: LLM Response Clustering



```python
def generate_response(narrative_text, queryForLLM):
    context = narrative_text
    prompt = f"Context:\n{context}\n\nQuestion:\n{queryForLLM}"

    # Tokenize prompt
    prompt_tokens = tokenizer.encode(prompt, add_special_tokens=False)
    max_length = 512  # Adjust based on model's max input length

    if len(prompt_tokens) > max_length:
        # Truncate context intelligently
        available_length = max_length - len(tokenizer.encode(f"Question:\n{queryForLLM}", add_special_tokens=False)) - 2
        # Re-tokenize the context to fit into available_length
        context_tokens = tokenizer.encode(context, add_special_tokens=False)
        truncated_context_tokens = context_tokens[:available_length]
        truncated_context = tokenizer.decode(truncated_context_tokens)
        prompt = f"Context:\n{truncated_context}\n\nQuestion:\n{queryForLLM}"

    # Generate the response
    output = pipe(
        prompt,
        max_new_tokens=512,  # Adjust as needed
    )
    return output[0]["generated_text"]
```

Code Block 02

Its workflow is described as follows. Narratives from law enforcement and coroner/medical examiner reports were combined and processed using the **LLama-3.2-3B-Instruct Model** to extract insights from unstructured inputs efficiently (Code Block 2). Text variables underwent cleaning to improve readability and consistency before being transformed into semantic embeddings using the **SBERT model (all-MiniLM-L6-v2)**. We employed **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) for clustering embeddings, leveraging its ability to identify clusters of arbitrary shapes and detect outliers. The optimal epsilon (**eps**) parameter for DBSCAN was determined through k-Nearest Neighbors distance analysis, ensuring robust cluster formation. Noise points were grouped separately, and the output included 102 cluster files and a noise file with 3,758 variable-related keywords, all prepared for expert validation.

Pathway 03: Combined pathway

We integrated LLM Response Clustering and Topic Modeling to develop this unified pathway. Through the help of expert knowledge, we extracted the related word that came from the topic modeling pathway and used it in chunk creation in similarity search and finally presented it then to the LLM model (Figure 04). These chunks were converted into high-dimensional embeddings capturing semantic meaning. A **FAISS**(Facebook AI Similarity Search) index was built for efficient semantic search, Narratives were concatenated into a single text for each record and tokenized into manageable chunks based on the maximum token length of the embedding model (**sentence-transformers/all-MiniLM-L6-** enabling query embeddings to retrieve the most relevant text chunks (Code Block 3).

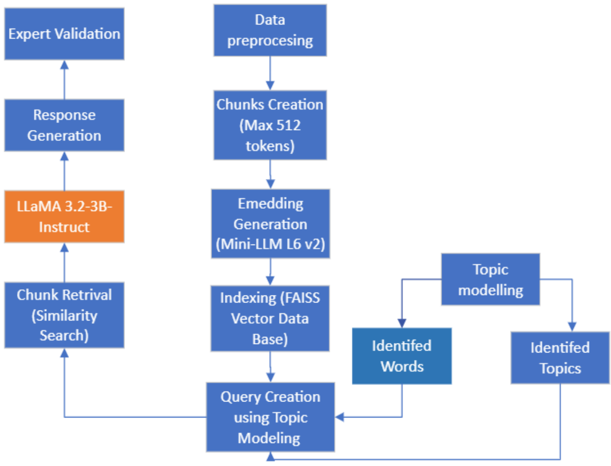
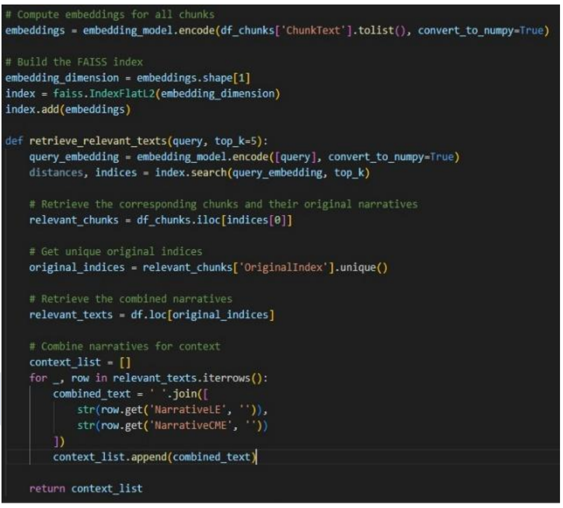

Expert Validation
Response Generation
LLaMA 3.2-3B-Instruct
Chunk Retrival (Similarity Search)
Data preprocesing
Chunks Creation (Max 512 tokens)
Emedding Generation (Mini-LLM L6 v2)
Indexing (FAISS Vector Data Base)
Query Creation using Topic Modeling
Topic modelling
Identifed Words
Identifed Topics

Figure 04: Topic Modeling + LLM Pathway



```
# Compute embeddings for all chunks
embeddings = embedding_model.encode(df_chunks['ChunkText'].tolist(), convert_to_numpy=True)

# Build the FAISS index
embedding_dimension = embeddings.shape[1]
index = faiss.IndexFlatL2(embedding_dimension)
index.add(embeddings)

def retrieve_relevant_texts(query, top_k=5):
    query_embedding = embedding_model.encode([query], convert_to_numpy=True)
    distances, indices = index.search(query_embedding, top_k)

    # Retrieve the corresponding chunks and their original narratives
    relevant_chunks = df_chunks.iloc[indices[0]]

    # Get unique original indices
    original_indices = relevant_chunks['OriginalIndex'].unique()

    # Retrieve the combined narratives
    relevant_texts = df.loc[original_indices]

    # Combine narratives for context
    context_list = []
    for _, row in relevant_texts.iterrows():
        combined_text = ' '.join([
            str(row.get('NarrativeLE', '')),
            str(row.get('NarrativeCME', ''))
        ])
        context_list.append(combined_text)

    return context_list
```

Code Block 03

Retrieved chunks were mapped back to their original narratives and combined into a coherent context. This system efficiently retrieves semantically relevant information for tasks like question answering or summarization. The output included 16 text files with identified variables, forwarded for expert validation.

Results

Through our comprehensive methodology, we successfully identified 45 novel variables related to suicide, critically examined by expert review. The identified variables represent important mental health determinants critical for understanding suicide risk and broader mental health contexts. These variables span across mental health domains, including prison-related contexts(2), sexual orientation (3), social media behaviours (4), substance use and suicidal ideation (5), physical health issues (6), social support dynamics, and social isolation (7).

Just to present these new variables in a user-friendly manner, we thought of categorizing these variables into two distinct groups: Category A comprises variables not directly mappable to the existing NVDR Manual categorization, while Category B includes variables that, although innovative, can be associated with existing NVDR Manual categories. To enhance clarity and utility, the new variables are presented in ***bold italics*** and linked to their most relevant NVDR category.

A. New variables

1. Family History: ***"Family history of mental illness", "Family history of suicidal ideations"***
2. Environmental factors: ***"Warning signs in the environment"***
3. Lack of support: "***Delayed response from emergency services"***
4. Appearance: ***"Weight loss"***
5. Relationships: ***"Social withdrawal", "Social isolation",*** "***Covid restrictions"***

6. Online behaviour / social media usage: **"suicidal intent on their cellphone"** / **"Searching for *information on how to harm themselves online*"** / ***"Addiction to Games."***
7. Education: ***"Academic pressure"***

B. Variable related to Existing NVDR Manual Variables

1) NVDR 3.1.20 (Military Services): ***"Discharged from the military",***
2) NVDR 3.1.6, 3.1.19,3.1.7 **(**Related to Sexuality)**:** ***"Questioning of their sexuality", "Prescribed hormone medication", "Sexual orientation and identity".***
3) NVDR 3.2.5 (Current Occupation): ***"Financials struggle."***
4) NVDR 4.3.3 **(**Prison Life): ***"Recent changes in the environment", "Fear of rejection by Family/Community"***, ***"Relationship with other inmates".***
5) NVDR 5.3, 5.3.5, (Substance Abuse): ***"Self-Medication", "Previous history of addiction relapse", "Drug paraphernalia"***.
6) NVDR 5.4.8 (Relationship): ***"Previous relationship issues"***
7) NVDR 5.5.1 (Crime): ***"Child pornography investigation"***
8) NVDR 5.7.4(Suicidal Ideation): ***"Past and wrote letters/Notes on laptop/*** **diary entries*"***, ***"Suicidal drawings"***,
8. NVDR 5.7.12 (Physical Health Problems): ***"Lack of access to treatment", "Migraine", "Lack of immediate medical attention",*** **"Noncompliance with medication"**
9) NVDR 6.3.10(Access to Firearms): ***"Presence of firearms in the residence", "Decedents access to firearms", "Gun safety knowledge", "Public access to firearms".***
10) NVDR 10.13 (Previous Arrests): ***"Police investigation", "Previous involvement with law enforcement".***
11) NVDR 10.14.9 **(**Mental Health Problems): ***"Prepartum depression", "Mental health treatment adherence".***
12) NVDR 11 (School): **"Switching schools"**
13) NVDR 11.3.3(Bullying): **"*Cyberbullying"***.
14) NVDR 12.4.4 (Availability Mental Health Services): ***"Availability of support systems", "Presence of School counselor".***

Conclusion

Through our innovative three-pathway pipeline, combining unsupervised machine learning techniques with expert medical validation, we successfully extracted a comprehensive set of novel variables from suicide narrative texts in the NVDRS dataset. The validated variables, curated by medical professionals, contribute valuable insights into suicide prevention strategies and future mental health research.

Further, we found that the application of domain-specific transformer models like ClinicalBERT, LegalBERT, and BioBERT was less effective in generating embeddings due to the lack of domain-specific terminology in the narratives, and we abandoned that approach.

Our data pipeline, which combines Topic Modeling and LLM approaches, demonstrates significant potential for adaptation across various healthcare use cases. This methodology allows users to customize variable extractions according to their specific requirements. Such flexibility makes it particularly valuable for developing initial frameworks or guidelines in healthcare settings where structured data extraction from narrative text is needed.

References

1. Akata Z, Balliet D, Rijke Md, Dignum F, Dignum V, Eiben G, et al. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. Computer. 2020;53(8):18-28.
2. Fazel S, Ramesh T, Hawton K. Suicide in prisons: an international study of prevalence and contributory factors. Lancet Psychiatry. 2017;4(12):946-52.
3. Russell ST, Joyner K. Adolescent sexual orientation and suicide risk: evidence from a national study. Am J Public Health. 2001;91(8):1276-81.
4. Maurya C, Muhammad T, Dhillon P, Maurya P. The effects of cyberbullying victimization on depression and suicidal ideation among adolescents and young adults: a three year cohort study from India. BMC Psychiatry. 2022;22(1):599.
5. Onaemo VN, Fawehinmi TO, D'Arcy C. Risk of suicide ideation in comorbid substance use disorder and major depression. PLoS One. 2022;17(12):e0265287.
6. Ahmedani BK, Peterson EL, Hu Y, Rossom RC, Lynch F, Lu CY, et al. Major Physical Health Conditions and Risk of Suicide. American Journal of Preventive Medicine. 2017;53(3):308-15.
7. Silva C, McGovern C, Gomez S, Beale E, Overholser J, Ridley J. Can I count on you? Social support, depression and suicide risk. Clinical Psychology & Psychotherapy. 2023;30(6):1407-15.