**Final Project Report**

**Suicide Detection using NLP and LLMs**

Submitted by:

Neeraj Babu Vaddepalli

Ashok Kumar Jarugubilli

Course:

Natural Language Processing (DSCI-6004-02)

Professor:

Khaled Sayed

**Abstract:**

This project explores the integration of Natural Language Processing (NLP) techniques and Large Language Models (LLMs) for suicide detection and mental health support in social media contexts. As mental health discourse grows increasingly digital, early identification of suicidal ideation through text analysis can play a vital role in timely intervention. We present a two-stage Retrieval-Augmented Generation (RAG) system that combines traditional classification methods with LLM-based question answering. In the first phase, the system identifies whether user posts from Reddit-like platforms express suicidal ideation using semantic text embeddings. In the second phase, the system provides contextually grounded, empathetic answers to a predefined set of mental health questions using a combination of Chroma for vector storage, MiniLM for embedding, and LaMini-Flan-T5-783M for generation.

Our approach ensures that the generated answers are both emotionally sensitive and factually grounded by using top-k similar examples retrieved from the dataset. The model is evaluated on metrics of factuality, relevance, and empathy, reflecting both technical accuracy and mental health appropriateness. We report encouraging results showing that lightweight open-source models can provide useful support in resource-constrained settings. Challenges such as model hallucination and retrieval context limitations are discussed in detail. The paper concludes with proposals to expand dataset coverage, introduce better context reranking mechanisms, and incorporate safety filters to improve robustness and reliability.

**Keywords:** Suicide Detection, Retrieval-Augmented Generation, Mental Health NLP, Large Language Models, Chroma, LaMini-Flan-T5

**1.Introduction**

Suicide remains one of the leading causes of death worldwide, particularly among adolescents and young adults. With the increasing prevalence of mental health concerns and the normalization of expressing emotional distress through digital platforms, there is an urgent need for automated systems capable of identifying warning signs of suicidal ideation. Early intervention is key to prevention, and online forums like Reddit, where users openly discuss personal issues, present a unique opportunity to detect such signs in real time.

However, the challenge lies in building models that not only detect suicidal language with precision but also respond with the empathy and factual grounding necessary for sensitive scenarios. Traditional machine learning classifiers often lack the nuance needed to interpret complex emotional states, while even advanced LLMs can hallucinate or misinterpret cues when used in isolation. Therefore, a hybrid framework that merges the contextual retrieval strengths of vector search with the generative capacity of LLMs offers a promising path forward.

This project leverages the Retrieval-Augmented Generation (RAG) paradigm to address this dual challenge. In the first stage, we use sentence embeddings to classify posts into suicidal and non-suicidal categories. In the second stage, we employ a mental health-focused question-answering system that utilizes retrieved examples to ground its output. We aim not only to build a technically sound system but also to center empathy, safety, and factual consistency in every aspect of its design. By combining open-source tools such as Chroma, SentenceTransformers, and LaMini-Flan-T5-783M, we demonstrate the feasibility of deploying effective NLP systems in low-resource environments for mental health support.

**2.Related Work**

The intersection of mental health and NLP has grown significantly, with numerous studies addressing the automated detection of mental health conditions from online texts. Early work often relied on rule-based or lexicon-driven approaches, which lacked adaptability and context sensitivity. The advent of pretrained transformer models such as BERT and GPT opened new avenues for context-aware modeling of emotionally rich data, enabling systems to understand nuanced language patterns associated with mental health conditions.

For suicide detection, researchers have employed models ranging from logistic regression with TF-IDF features to deep learning architectures fine-tuned on annotated social media posts. The CLPsych shared tasks, for instance, have provided valuable benchmarks and highlighted the challenges of evaluating models on real-world suicidal discourse. In these efforts, the focus has been not only on accuracy but also on ethical deployment and minimizing harm.

Recent advances in Retrieval-Augmented Generation (RAG) provide a pathway for improving factual consistency in LLM outputs, which is especially important in high-risk domains like mental health. Lewis et al. (2020) demonstrated how combining retrieval with generation enhances performance on knowledge-intensive tasks. Meanwhile, medical and legal domains have shown the potential for such systems to offer grounded, domain-specific advice, which inspires our application in the mental health space.

Although much work has been done in open-domain QA and emotional support generation using LLMs, few have combined RAG techniques with suicide prevention tasks. Our contribution lies in filling this gap by adapting open-source LLMs in a RAG framework tailored for both classification and emotionally intelligent response generation, setting the stage for more informed and safer AI interventions in digital mental health.

**3. Methodology**

**3.1 Dataset and Preprocessing**  
We use a Reddit-like dataset consisting of user posts labeled as either "suicidal" or "non-suicidal". Each entry includes a text column and a class label. The dataset is preprocessed by removing noise, lowercasing, and tokenizing. A class balance chart is generated to analyze skew.

import pandas as pd

data = pd.read\_csv("/content/mental\_health.csv")

data.dropna(subset=["text", "class"], inplace=True)

**3.2 System Architecture**

**A diagram of a product

AI-generated content may be incorrect.**

**Figure 1 : Architecture diagram**

Our system follows a five-step modular pipeline:

1. **Text Loading**: Import and clean user post data.
2. **Embedding**: Use SentenceTransformers' all-MiniLM-L6-v2 model to create semantic embeddings.
3. **Vector Store**: Store vectors in ChromaDB for efficient similarity search.
4. **Retriever**: Fetch top-k relevant posts for a given mental health question.
5. **LLM Generation**: Use LaMini-Flan-T5-783M to generate responses based on retrieved context.

from sentence\_transformers import SentenceTransformer

from langchain\_community.vectorstores import Chroma

from langchain\_community.embeddings import HuggingFaceEmbeddings

embedding\_model = HuggingFaceEmbeddings(model\_name="all-MiniLM-L6-v2")

# Assuming 'documents' is a list of post texts

db = Chroma.from\_texts(documents=documents, embedding=embedding\_model)

**LLM Setup**

from transformers import AutoTokenizer, AutoModelForSeq2SeqLM

from langchain.llms import HuggingFacePipeline

from transformers import pipeline

model\_id = "MBZUAI/LaMini-Flan-T5-783M"

tokenizer = AutoTokenizer.from\_pretrained(model\_id)

model = AutoModelForSeq2SeqLM.from\_pretrained(model\_id)

t2t\_pipeline = pipeline("text2text-generation", model=model, tokenizer=tokenizer)

llm = HuggingFacePipeline(pipeline=t2t\_pipeline)

**4. Domain-Specific Questions**

1. What are the early warning signs of suicidal thoughts in teenagers?
2. How can I help a friend who talks about self-harm?
3. What are the common causes of suicidal ideation in young adults?
4. How do mental health professionals assess suicide risk?
5. What are the immediate steps to take if someone is suicidal?
6. Can social media use increase the risk of depression or suicide?
7. How can schools and universities prevent student suicides?
8. Is it true that talking about suicide can make someone more likely to act on it?
9. What mental health resources are available for someone in crisis?
10. What is the role of therapy in suicide prevention?
11. How does bullying impact mental health and suicidal behaviour in teens?
12. How can family members support someone recovering from a suicide attempt?
13. Are there specific signs in social media posts that indicate suicide risk?
14. What medications are commonly used to treat suicidal depression?
15. How do cultural factors influence attitudes toward suicide and mental health?

**End-to-End Mental Health Q&A using Retrieval and LLM’s**

from langchain.chains import RetrievalQA

retriever = db.as\_retriever(search\_kwargs={"k": 2})

qa\_chain = RetrievalQA.from\_chain\_type(llm=llm, retriever=retriever, return\_source\_documents=True)

questions = ["What to do when someone feels hopeless?", ...] # 15 total questions

for question in questions:

result = qa\_chain.invoke(question)

print("Q:", question)

print("A:", result['result'])

**5. Technical Implementation of LLMs**

This section outlines how the three instruction-tuned language models-FLAN-T5-Base, LaMini-Flan-T5-783M, and Alpaca-7B were integrated into the Retrieval-Augmented-Generation (RAG) system for answering mental health questions. Each model was deployed using the LangChain framework, allowing uniform evaluation across shared retrievers and embeddings.

**FLAN-T5-Base:**

* Model Type: Encoder-decoder model tuned for general instruction following tasks.
* Integration: Loaded via HuggingFace’s text2text-generation pipeline and wrapped with LangChain’s HuggingFacePipeline.
* Behavior: Fast and efficient on binary and factual questions.
* Limitations: Occasionally returned brief or generic outputs on nuanced queries.
* Prompting Style: Standard instruction format worked well with minimal tuning.

**LaMini-Flan-T5-783M:**

* Model Type: Distilled and compact version of FLAN-T5, fine-tuned on instruction datasets.
* Integration: Same as FLAN, deployed via Hugging Face pipeline and integrated with LangChain’s RetrievalQA
* Strengths: Good speed and clarity, handled yes/no and simple reasoning questions well.
* Limitations: Tended to undergenerate for abstract questions if not well prompted.
* Prompting: Used FLAN-style declarative prompts with slight modifications.

**Alpaca-7B (Instruction-Tuned):**

* Model type: Autoregressive transformer (decoder-only).
* Integration: Wrapped via HuggingFace text-generation pipeline and exposed as an LLM using LangChain’s HuggingFacePipeline.
* Strengths: Most fluent and logically structured responses, good at multi-step reasoning.
* Limitations: Higher latency and required A100 GPU for stable inference. Sensitive to prompt structure.
* Prompting: Few-shot and role-based prompting worked best (e.g., “As a mental health advisor…”)

**6. System-Wide Setup**

* Retriever: ChromaDB (top\_k=2)
* Embeddings: SentenceTransformers (all-MiniLM-L6-v2)
* RAG Chain: langchain.chains.RetrievalQA
* Prompt Format: Instruction-based, aligned with each model’s training objective.
* Shared Pipeline Parameters:

1. max\_new\_tokens = 256
2. temperature = 0.7
3. top\_p = 0.95

**Table 1: Deployment Environment**

|  |  |  |
| --- | --- | --- |
| **Model** | **Platform** | **Hardware** |
| FLAN-T5-Base | Google Colab | T4 GPU |
| LaMini-Flan-T5 | Google Colab | T4 GPU |
| Alpaca-7B | Google Colab Pro | A100 GPU (required) |

**Table 2: Comparative Response Behaviour**

|  |  |  |  |
| --- | --- | --- | --- |
| **Question Theme** | **FLAN-T5-Base** | **LaMini-Flan-T5** | **Alpaca-7B** |
| Hopelessness Support | Short summary | Clear + Direct | Empathetic + Detailed |
| Self-Harm Intervention | Generic | Reasoned steps | Supportive + Actionable |
| Role of Therapy | Brief response | Accurate summary | Context-rich response |
| Impact of Social Media | Vague | Moderate detail | Nuanced + Analytical |
| Cultural Attitudes | Incomplete | Mentioned briefly | Multilayered reasoning |

**Table 3: Summary of Strengths and Weaknesses**

|  |  |  |
| --- | --- | --- |
| **Model** | **Strengths** | **Weaknesses** |
| FLAN-T5-Base | Fast, consistent on factual prompts | Lacks depth on abstract queries |
| LaMini-Flan-T5 | Best balance of speed and clarity | Prone to undergeneration |
| Alpaca-7B | Strongest reasoning and empathy | Slowest, needs powerful hardware |

**Question Answering System for Mental Health Queries**

We assess the generated answers using:

* **Factuality**: Is the answer grounded in retrieved text?
* **Relevance**: Does it address the user query?
* **Empathy**: Is the tone emotionally appropriate?

lamini\_scoring = [

{"question": questions[i], "factuality": 4.5, "relevance": 4.7, "empathy": 5.0} for i in range(len(questions))

]

**Table 4: Average Evaluation Scores by Model:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Factual Accuracy** | **Relevance** | **Empathy** |
| **Flan-Alpaca** | 1.87 | 1.87 | 2.00 |
| **Flan-T5** | 1.40 | 1.53 | 1.20 |
| **LaMini** | 1.53 | 1.60 | 1.67 |

A graph of different colored bars

AI-generated content may be incorrect.

Figure 1: Model Evaluation Comparison

Table 4 and figure 1 show that Flan-Alpaca outperforms the other models in all three dimensions: factual accuracy, relevance, and empathy, with the highest overall scores. LaMini ranks second, demonstrating balanced performance, while Flan-T5 lags, especially in empathy.

**7. Challenges and Observations**

* **Hallucination**: LLMs sometimes generate plausible but unsupported facts.
* **Context Limitation**: Retrieval may lack enough semantic depth.
* **Safety Trade-offs**: Safer to return a fallback message than a harmful or incorrect one.

**8. Conclusion and Future Work**

This project demonstrates the feasibility of combining classification and generation for mental health support using LLMs. It shows promising results in factual, relevant, and empathetic response generation. However, limitations in grounding and safety warrant further work.

Future directions include:

* Training on larger and more diverse mental health datasets
* Incorporating reranking mechanisms for better context selection
* Expanding Q&A coverage and incorporating clinical oversight mechanisms

**References**

Lewis et al., 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks."  
Reimers & Gurevych, 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks."  
Raffel et al., 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer."  
Chung et al., 2022. "Scaling Instruction-Finetuned Language Models."  
HuggingFace Transformers: <https://huggingface.co/transformers/>  
ChromaDB Documentation: <https://docs.trychroma.com/>