Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Advanced Sciences

Department of Mathematics

Winter Semester 2024-25

MDT6099 - Masters Thesis

Final Review

NAME : PRAVEEN. T **REG.NO** : 23MDT0051

Mental health prediction using Transcript Data

UNDER THE GUIDANCE OF: Dr. MANIMARAN A

Table of Content

PROBLEM STATEMENT

- Mental health disorders such as depression, anxiety, and PTSD (Post-Traumatic Stress Disorder) often go undetected due to manual diagnosis methods and limited clinical resources.

- Traditional methods are time-consuming, inconsistent, and not scalable for early intervention.

- Vast amounts of therapy session transcript data remain underutilized in clinical decision-making.

For this problem, we need an intelligent system that can analyze unstructured text to accurately predict mental health conditions using advanced NLP and deep learning techniques.

OBJECTIVE

- The objective of this project is to develop an advanced mental health detection system using NLP techniques to accurately analyze textual data from therapy transcript data.

- By utilizing BERT for contextual understanding and fine-tuning the model with the AdamW optimizer, the project aims to enhance the accuracy of mental health prediction.

LITERATURE SURVEY

1. Psychiatry transcript annotation: Process study and improvements

This study addresses the lack of a standardized process for mental health transcript annotation, essential for training reliable AI models. Three clinicians and five non-expert subjects annotated transcripts in two phases before and after training. Results showed improved inter-rater reliability and accuracy after training, highlighting the importance of clear labeling and annotator preparation. These findings support the development of efficient data collection methods to aid psychiatrists and improve machine learning applications in mental health.

2. Analysis of Therapy Transcripts using Natural Language Processing

Mental health is essential for overall well-being, and early detection of disorders is crucial due to their long-term effects. With over 450 million people currently affected, NLP offers a powerful way to analyze therapy transcripts for early signs of mental health issues. Our system classifies responses into 'Early signs of depression' and 'Serious after-effects of prolonged depression' using classifiers like Naïve Bayes, SVM, and Logistic Regression, along with TF-IDF and Count Vectorization. This tool aids both patients and therapists by enabling early diagnosis, large-scale data analysis, and more effective therapy interventions.

3. Classification and analysis of text transcription from Thai depression assessment tasks among patients with depression

Depression is a major health concern in Thailand, worsened by limited mental health services. This study evaluates XLM-RoBERTa, a multilingual NLP model, for depression classification from Thai speech transcripts. Using three key assessment questions, the model achieved 90% accuracy. Analysis revealed that depressed individuals used negative words like ‘sad’ and ‘suicide,’ while controls used neutral/positive terms. Findings suggest that brief text-based screening can aid early depression detection, reducing healthcare burden.

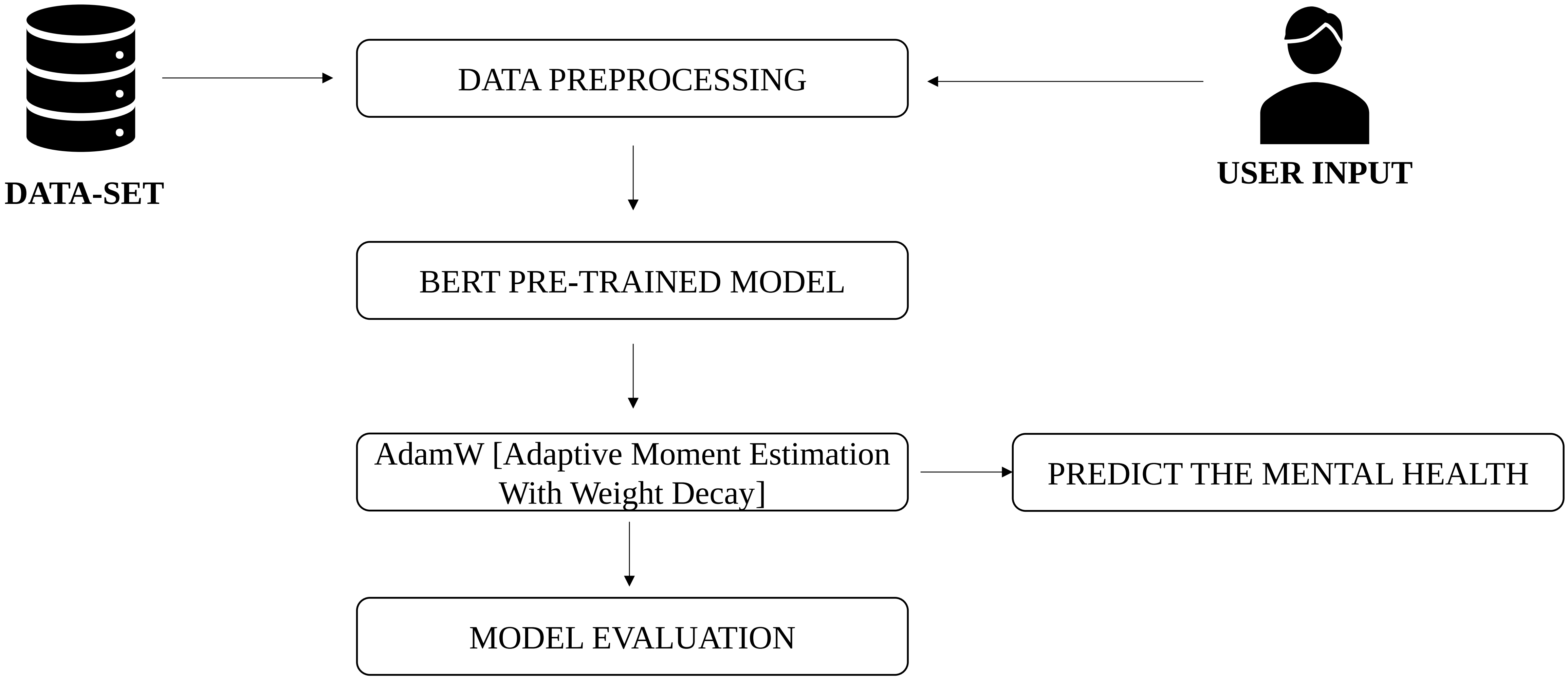
4. Automated clinical transcription for behavioral health clinicians

Mental health disorders are common but often go untreated due to limited resources and clinicians. Electronic Health Records (EHRs) aid documentation but are time-consuming, leading to clinician burnout. This study develops an automated clinical transcription tool to extract key information from patient-provider conversations and generate clinical notes. Using 65 simulated transcripts, a fine-tuned transformer model achieved F1=0.94 for extraction and a rule-based module synthesized notes, reducing documentation time by 70-80%. This work enhances behavioral healthcare efficiency and NLP applications in clinical settings.

DATA SET

	A	B	C	D	E	F	G	H	I
1	questionID	questionTitle	questionText	questionUrl	topics	therapistName	therapistUrl	answerText	upvote
2	5566fab2a64752d7	Escalating disagree	My wife and	https://counselchat.cc	Family Conflict	Kristi King-Morgan, L	https://counselc	<p>What you are describing	
3	5566f94fa64752d7	I'm addicted to sm	I'm planning to	https://counselchat.cc	Substance Abus	Rebecca Duellman	https://counselc	<p>Hi. Good for you in plan	
4	5567d26887a1cc0c	Keeping secrets fro	I have secrets in	https://counselchat.cc	Family Conflict	Jeevna Bajaj	https://counselc	<p>It sounds like keeping th	
5	556bed15c969ba5	The Underlying Cau	I am extremely p	https://counselchat.cc	Behavioral Char	Rebecca Duellman	https://counselc	<p>Hi there. It's great you a	
6	556ba115c969ba5	Can I control anxiet	I had a head injur	https://counselchat.cc	Anxiety	Rebecca Duellman	https://counselc	<p>You didn't say what or h	
7	556b6940c969ba5	How do I break an	I want a secure	https://counselchat.cc	Relationship Dis	Kristi King-Morgan, L	https://counselc	<p>It is a good thing that yc	
8	556bec8cc969ba58	I have anger issues	I easily recognize	https://counselchat.cc	Anger Managen	Kristi King-Morgan, L	https://counselc	<p>I suggest that you work	
9	5566f9a2a64752d7	Iâ€™ve suffered frc	It takes me a	https://counselchat.cc	Sleep Improven	Danielle Alvarez	https://counselc	<p>First of all, exercise is alv	
10	5570b7fea03de6c3	Unethical Therapy	What do you do v	https://counselchat.cc	Professional Eth	Kristi King-Morgan, L	https://counselc	<p>I will admit I am confuse	
11	556bf606c969ba58	My friends accusin	They're calling m	https://counselchat.cc	Social Relations	Danielle Alvarez	https://counselc	<p>It sounds like your confu	
12	55711873a03de6c3	About a year ago I	Cheating is	https://counselchat.cc	Relationships,M	Danielle Alvarez	https://counselc	<p>First of all, my heart goe	
13	55717c13a03de6c3	Sleeping, Anger an	I have a lot of iss	https://counselchat.cc	Anxiety,Anger M	Danielle Alvarez	https://counselc	<p>It sounds as if you may	
14	5571cff7a03de6c3	I'm losing my husb	I have no sex	https://counselchat.cc	Marriage,Intima	Danielle Alvarez	https://counselc	<p>Iâ€™m sorry to hear abc	
15	55717c13a03de6c3	Sleeping, Anger an	I have a lot of iss	https://counselchat.cc	Anxiety,Anger M	Keisha Helms	https://counselc	<p>Hi there. I have to comm	
16	557136aaa03de6c3	I need help of letting go of a man wh		https://counselchat.cc	Relationships	Danielle Alvarez	https://counselc	<p>It is incredibly hard to le	

METHODOLOGY



1. DATA PREPROCESSING:

In this process, I analyze data and use some preprocessing methods to clean the data set process for BERT embedding and process the BERT pre-trained model.

Steps:

Handle NaN or non-string values → Remove HTML tags → Remove special characters → Convert to lowercase →

Remove stop words, punctuation, and lemmatize → labeling based on the category

2. BERT PRE-TRAINED MODEL:

- After completing the **"Data Preprocessing,"** I load a pre-trained BERT, which is a ready-made **BERT model** for text classification and adapts BERT to understand specific **dataset labels**.
- The model assigns categories based on **unique labels** in the dataset.
- Tokenization used the BERT Tokenizer to convert text into numerical format for model processing.

3. AdamW [Adaptive Moment Estimation With Weight Decay] :

After **the BERT pre-train model** step, I define the optimizer, using the **AdamW optimizer** to adjust BERT's weights for better accuracy. With the help of Training Loop, I train the model over multiple epochs to improve predictions. Compute loss & update weights and calculate training loss and update model weights for improved learning.

4. MODEL EVALUATION:

With the help of Sklearn matrix the model evaluated and **classification report** is generated

Accuracy: 0.9507

	precision	recall	f1-score	support
0	0.88	0.97	0.93	234
1	0.98	0.92	0.95	155
2	0.98	0.95	0.97	62
3	0.97	1.00	0.99	38
4	0.98	0.95	0.97	436
5	0.95	0.95	0.95	39
6	1.00	1.00	1.00	26
7	0.82	1.00	0.90	9
8	0.92	1.00	0.96	120
9	0.95	0.92	0.94	363
accuracy			0.95	1482
macro avg	0.94	0.97	0.95	1482
weighted avg	0.95	0.95	0.95	1482

Accuracy: **0.95** → 95% of all predictions were correct.

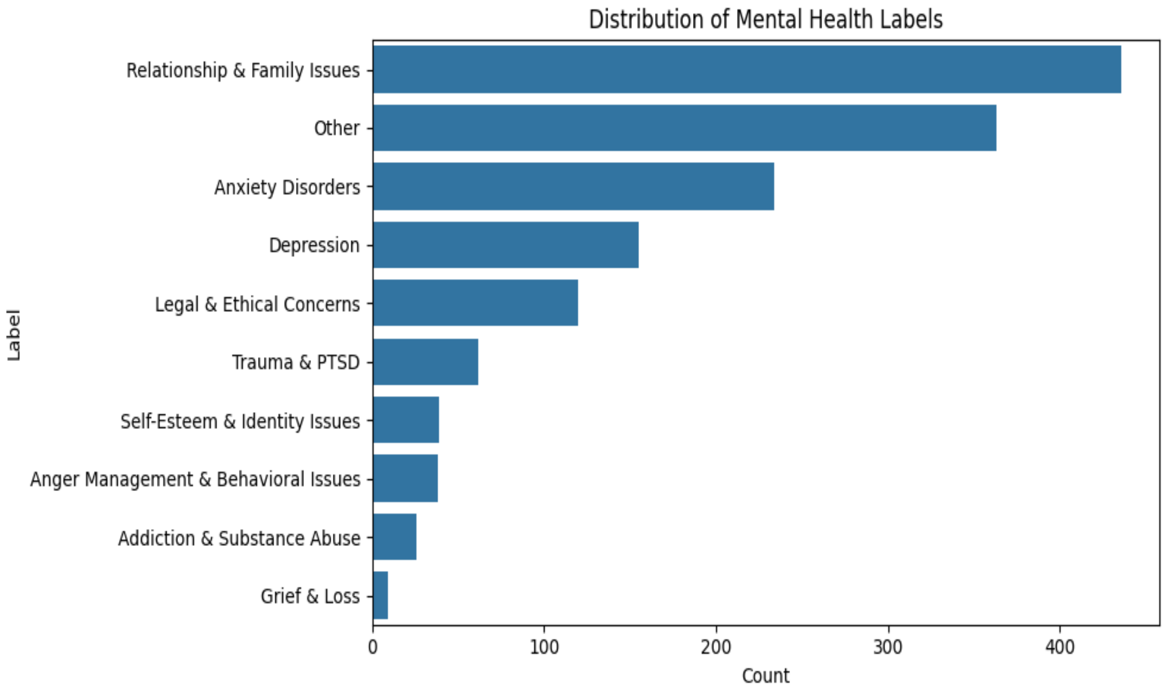
Macro Average (unweighted mean across all classes):

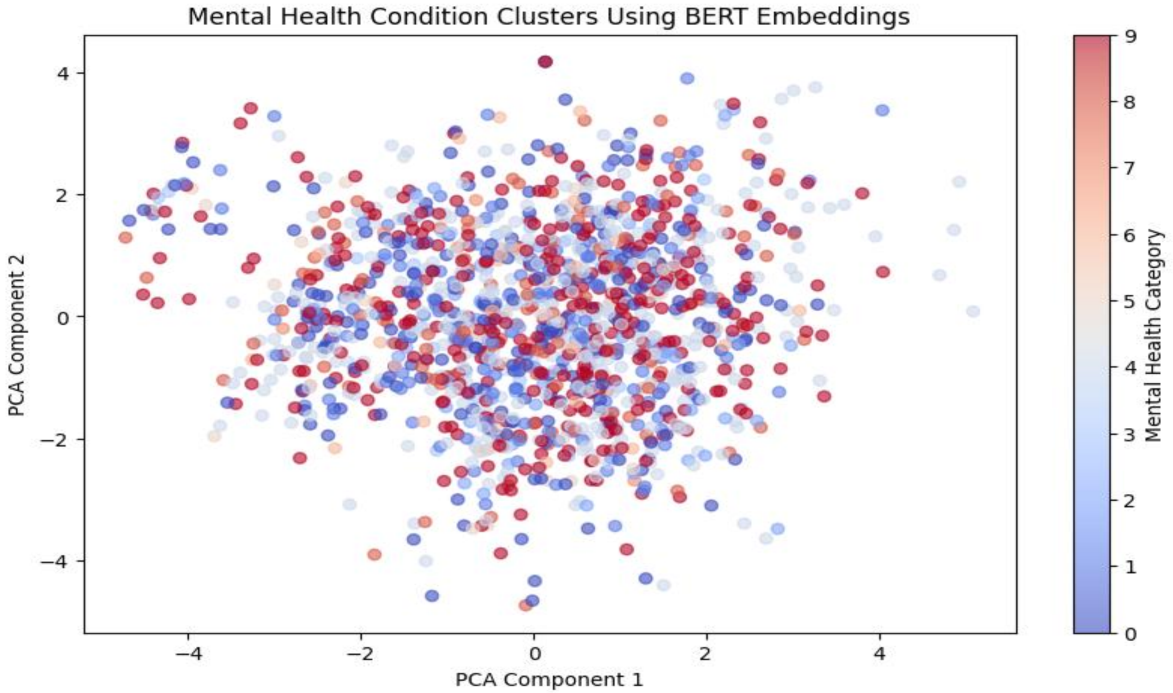
Precision: **0.94**
Recall: **0.97**
F1-score: **0.95**

Weighted Average (weighted by support count):

Precision: **0.95**
Recall: **0.95**
F1-score: **0.95**

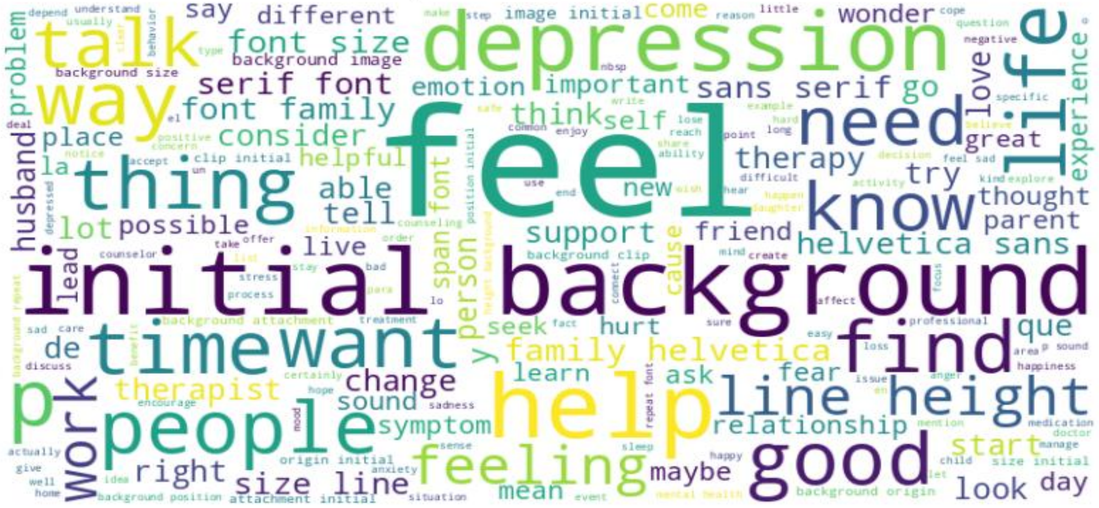
EXPLORATORY DATA ANALYSIS





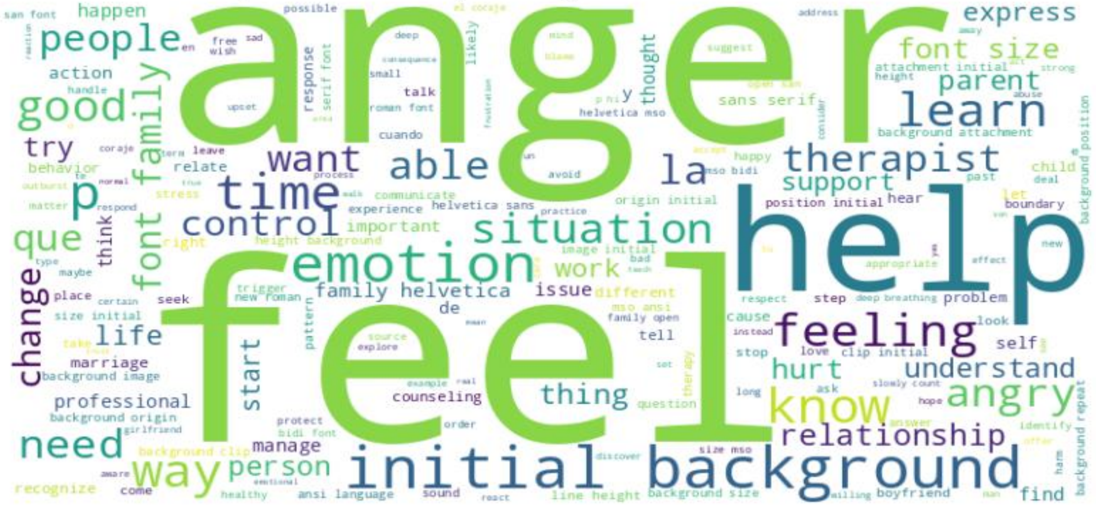
Most Frequent Words in Anxiety Disorders Texts

Most Frequent Words in Depression Texts



Most Frequent Words in Trauma & PTSD Texts

Most Frequent Words in Anger Management & Behavioral Issues Texts



RESULTS

1.

```
please free to share of your life: there is full of relationship issues between me and lover, she not taking with me for past three week.

 Predicted Mental Health Condition:
   @ Topic: Relationship & Family Issues (Label: 4)
   @ Confidence Score: 0.98
--------------------------------------------------------------------------

 Top 3 Predicted Mental Health Conditions:
   1. Relationship & Family Issues (Label: 4, Confidence Score: 0.98)
```

2.

```
please free to share of your life: i fail in exam, i am feeling not good

 Predicted Mental Health Condition:
   @ Topic: Anxiety Disorders (Label: 0)
   @ Confidence Score: 0.79
--------------------------------------------------------------------------

 Top 3 Predicted Mental Health Conditions:
   1. Anxiety Disorders (Label: 0, Confidence Score: 0.79)
```

CONCLUSION

This project successfully developed an **NLP-based mental health classification system** using **therapy session transcripts**. The system utilizing **BERT embeddings** and fine-tuned the model using the AdamW optimizer to improve performance. That helps to predict mental health concerns based on user queries and therapist responses.

FUTURE WORKS

- **Multilingual Support:** Enable the system to work in multiple languages for broader accessibility and inclusion.

- **Explainability:** Use SHAP(Shapley Additive Explanations) to help users and professionals understand how the model makes predictions.

- **Hybrid Modeling:** Combine BERT-based NLP with psychological knowledge graphs or rule-based logic for better contextual understanding.

- **Clinical Validation:** Collaborate with mental health experts to validate model predictions against actual clinical diagnoses.

- **Privacy-Preserving Techniques:** Implement method differential privacy to protect user data while maintaining accuracy.

- **Real-Time Chatbot:** Integrate the model into a live chatbot for instant mental health support and feedback.

REFERENCE

[1]. Tlachac, M. L., Toto, E., Lovering, J., Kayastha, R., Taurich, N., & Rundensteiner, E. Emu: Early mental health uncovering framework and dataset. In *2021 20th IEEE international conference on machine learning and applications (ICMLA)*, 2021, December, pp. 1311-1318 https://ieeexplore.ieee.org/abstract/document/9680143

[2]. Verma, P., & Shakya, M. *Machine learning model for predicting Major Depressive Disorder using RNA-Seq data: Optimization of classification approach. Cogn Neurodyn 2022; 16 (2): 443-53.*https://link.springer.com/article/10.1007/s11571-021-09724-8

[3]. Munthuli, A., Pooprasert, P., Klangpornkun, N., Phienphanich, P., Onsuwan, C., Jaisin, K., ... & Tantibundhit, C. (2023). Classification and analysis of text transcription from Thai depression assessment tasks among patients with depression. *PLoS one*, *18*(3), e0283095.https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0283095

[4]. Kazi, N. H. (2022). *Automated clinical transcription for behavioral health clinicians* (Doctoral dissertation, Montana State University-Bozeman, College of Engineering). https://scholarworks.montana.edu/items/1ccf24ec-86d7-46a4-ab1d-ba67041e6d96

[5]. Kaynak, E. B., & Dibeklioğlu, H. Systematic analysis of speech transcription modeling for reliable assessment of depression severity. *Sakarya University Journal of Computer and Information Sciences*,2024, vol *7*(1), pp. 77-91. http://saucis.sakarya.edu.tr/en/pub/issue/84270/1381522