Nile University

School of Information Technology and Computer Science

Program of Computer Science

Chezlong: Arabic Chatbot for Mental Health Support

CSCI495 Senior Project II
Submitted in Partial Fulfilment of the Requirements
For the Bachelor’s Degree in Information Technology and Computer Science
Computer Science

Submitted by

Muhammad Helmy

1610114

Supervised by

Dr. Ghada Khoriba

Giza – Egypt

Fall 2024

Project Summary

Technology has opened up a new era of mental health care and information gathering, providing new means to access aid, connect to a peer counsellor, and track progress. Chezlong is an Arabic text-based chatbot powered with the most modern NLP techniques. Its aim is to give guidance for those who need a quick mental support. The development cycle is based on using GPT-3.5 Large Language Model for Retrieval Augmented Generation (RAG). RAG is used on Arabic psychological articles, originally scrapped from nafsy.net website. Cleaning, pre-processing, and topic modelling has been applied to prepare and explore the dataset. The chunked dataset is stored as vectors in Pinecone. Similarity search is done to retrieve query-related source knowledge from the dataset to answer a user query.

Keywords: Chatbot, Arabic, NLP, LLM, RAG, Psychology

Table of Contents

List of **Figures**

Chapter 1

Introduction

Chezlong is an Arabic and English text-based chatbot. Its aim is to help patients with mental health disorders by interacting with them in a humanly analogous and sympathetic way. Users can get insightful information easily and cheaply.

1.1. Background:

Chatbots are intelligent conversational computer systems designed to mimic human communication in order to provide automatic online guidance and help. They make use of approaches and algorithms from two AI domains: Natural Language Processing and Machine Learning. Chatbots typically accept natural language text as input and output the most relevant output to the user input sentence [1].

The history of chatbots dates back to the 1950's when Alan Turing proposed a method to determine a computer's level of intelligence. One of the first chatbots was ELIZA, which failed the Turing Test despite being able to make some users believe they were speaking to a real person. In 1972, Kenneth Colby produced PARRY, a chatbot that could impersonate a person suffering from paranoid schizophrenia. In 1995, A.L.I.C.E. was a language-processing bot that was widely used online, but it was unable to pass the Turing Test. Big tech companies began to use bots heavily throughout the decade, starting with Siri (2010), Google Now (2012), Alexa (2015), and Cortana (2015). These bots surprised users with their capabilities to perform a variety of tasks, such as playing music, running internet searches, and responding to voice requests [2][3]. However, ChatGPT is the current huge global surprise. ChatGPT, developed by OpenAI, went online in November 2022. It is based on the OpenAI GPT-3 family of large language models and has been tuned using both supervised and reinforcement learning techniques [4].

Chatbots are increasingly popular due to the increase in computational power. Recent developments in Artificial Intelligence and Natural Language Processing techniques have made them easier to implement, but human-chatbot interaction is not perfect. Chatbots can provide support in different fields as well as entertainment and companionship for the end user. Deep Learning algorithms have enabled the development of smart personal assistants, such as Amazon's Alexa, Apple's Siri, Google's Assistant, Microsoft's Cortana, and IBM's Watson. Chatbots have a personality, an Intelligent Quotient (IQ) and Empathy and Social Skills (EQ). They can also provide entertainment and assist with day-to-day tasks. The core implementation approaches to chatbots are Rule-Based and Artificial Intelligence with Transformers, first presented in the paper "Attention is all you need", as the state of the art in the deep learning algorithms [1]. Also, there are some other domain-specific chatbots like Chatbot System on the Teaching of Foreign Languages [15] and chatbot-based healthcare service with a knowledge base for cloud computing providing fast treatment in response to accidents and chronic diseases [16].

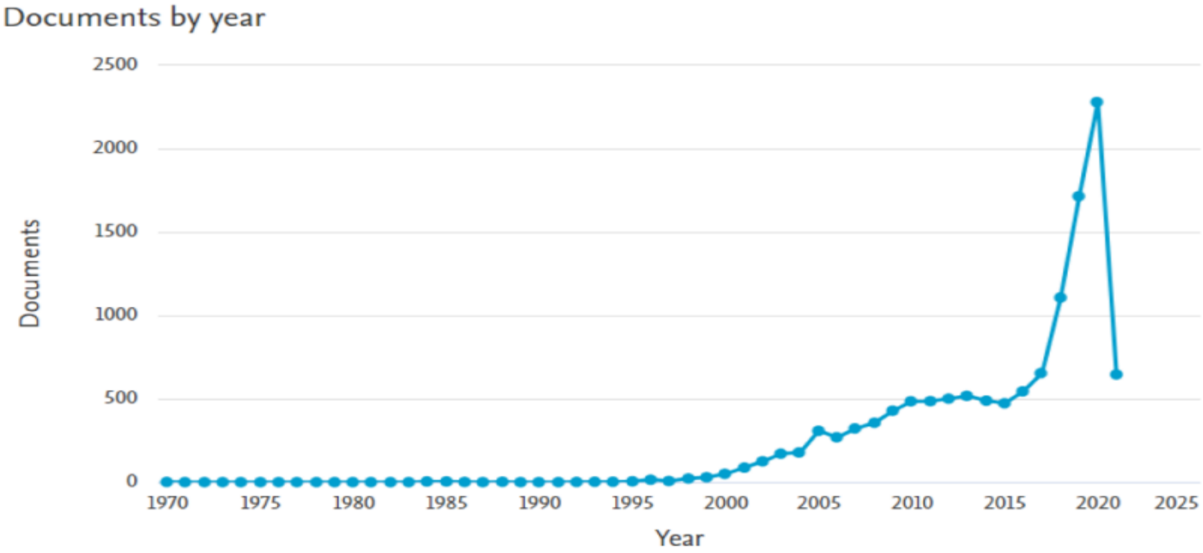
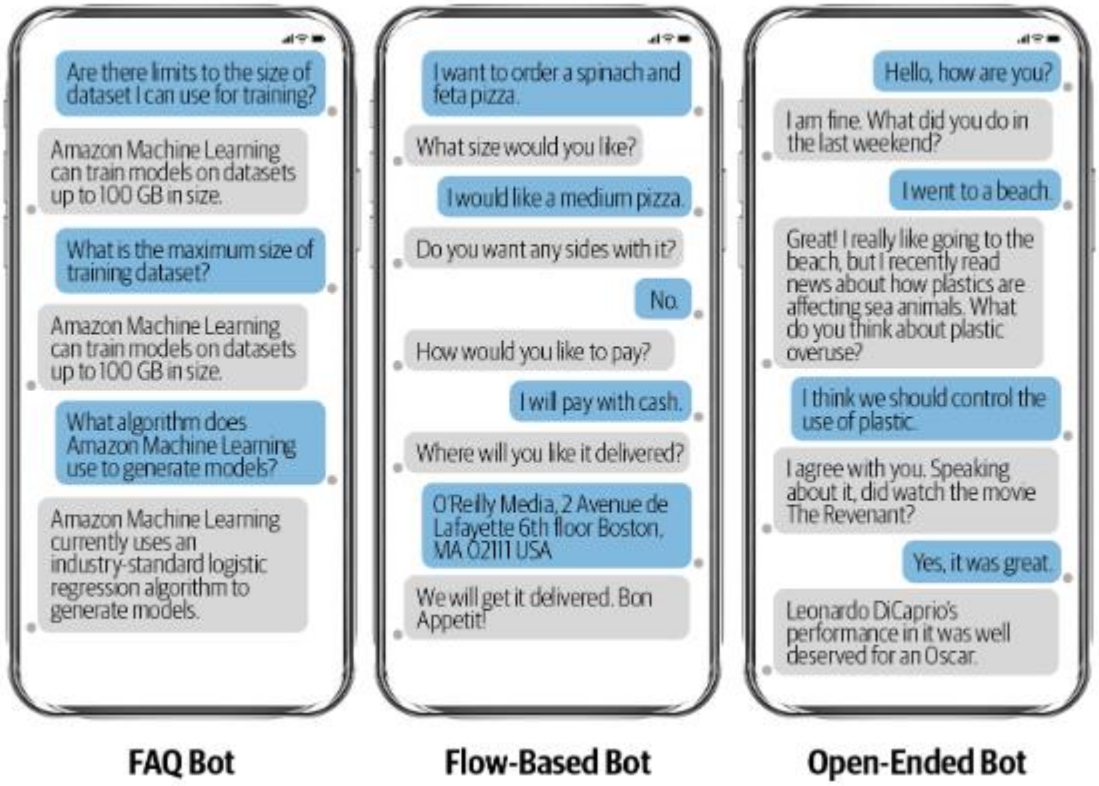


Figure (1): Search Results from Scopus, from 1970 to 2021 for the keywords “chatbot” or “conversational agents” or “conversation system” [1].

Taxonomy of chatbots is divided into two groups:

- Goal Oriented Dialogue: Specific use case where the end aim is known. Goal-oriented chatbots are domain-specific and require the system's integration of domain-specific knowledge, which limits the chatbot framework's capacity to be generalised and scaled. Examples of Goal Oriented Dialogue Chatbots are FAQ bots and flow-based bots (fig (8)).
- Chitchats: unstructured, open-domain conversations between humans that involve free-form, opinionated discussions about various topics. They are challenging due to the absence of objective goals and must generate coherent, on-topic, and factually correct responses to make the dialog more natural. The application of chitchat bots is futuristic but holds immense potential, such as using them to elicit useful but sensitive information in the case of a medical emergency or address loneliness and depression among teenagers and elderly people. Market-leader companies such as Amazon, Apple, and Google are investing heavily in building such bots for worldwide customers [21].

Chezlong is domain specific for mental health purposes, but it is basically a chitchat bot.



Are there limits to the size of dataset I can use for training?

Amazon Machine Learning can train models on datasets up to 100 GB in size.

What is the maximum size of training dataset?

Amazon Machine Learning can train models on datasets up to 100 GB in size.

What algorithm does Amazon Machine Learning use to generate models?

Amazon Machine Learning currently uses an industry-standard logistic regression algorithm to generate models.

I want to order a spinach and feta pizza.

What size would you like?

I would like a medium pizza.

Do you want any sides with it?

No.

How would you like to pay?

I will pay with cash.

Where will you like it delivered?

O'Reilly Media, 2 Avenue de Lafayette 6th floor Boston, MA 02111 USA

We will get it delivered. Bon Appetit!

Hello, how are you?

I am fine. What did you do in the last weekend?

I went to a beach.

Great! I really like going to the beach, but I recently read news about how plastics are affecting sea animals. What do you think about plastic overuse?

I think we should control the use of plastic.

I agree with you. Speaking about it, did watch the movie The Revenant?

Yes, it was great.

Leonardo DiCaprio's performance in it was well deserved for an Oscar.

FAQ Bot

Flow-Based Bot

Open-Ended Bot

Figure (2): Types of Chatbots.

1.2. Motivation:

Technology has opened up a new era of mental health care and information gathering, providing doctors and researchers with new means to access aid, connect to a peer counsellor, track progress, and gain a better understanding of mental health. New technology can be packaged into an app that can collect information on a user's behavior patterns and signal that assistance is required before a crisis occurs. The following are just a few benefits:

- Convenience: Treatment can be done anywhere and anytime, making it ideal for those with difficulty with in-person appointments.
- Anonymity: Clients can seek treatment without involving others.
- An introduction to care: Technology can help those who have avoided mental health care.
- Lower Cost: Apps are free or cost less than traditional care.
- Service to more people: Technology can help provide mental health services in remote areas or in times of crisis.
- Interest: Technology may be more appealing than traditional treatments, encouraging clients to continue therapy.
- 24-hour service
- Support: Technology can complement traditional therapy by reinforcing skills and providing support and monitoring.
- Objective data collection: Technology can collect and store data about location, movement, phone use, and other information.

However, Mental health community and software developers are focusing on addressing potential problems to ensure new apps provide benefits without harm:

- Effectiveness: Scientific evidence is essential to ensure technological interventions work as well as traditional methods.
- For whom and for what: Another concern is understanding if apps work for all people and for all mental health conditions.
- Privacy: App makers must ensure user privacy to protect sensitive personal information.
- Guidance: No industry-wide standards exist to assess the effectiveness.

- Regulation: The question of who should regulate mental health technology and data needs to be addressed.
- Overselling: Consumers may turn away from effective therapies if an app or program promises more than it delivers.

Some popular areas of app development include Self-Management Apps, Skill-Training Apps, Supported Care, Passive Symptom Tracking, and Data Collection [5]. Moreover, there are five technological innovations transformed the way mental health services were provided in 2020:

- Prescription Video Games: EndeavorRX is a prescription video game for kids with ADHD that challenges them to focus on multiple tasks simultaneously. It harnesses modern kids' natural habits to drive measurable improvements in their mental health.
- AI- And Smartphone-Assisted Therapy: AI-powered tools such as Woebot and Wysa can help patients practice CBT strategies and manage their symptoms between appointments. Teletherapy 2.0 combines inputs from multiple modalities to improve patient outcomes.
- VR For Mental Health: VR can help treat anxiety, depression, and post-traumatic stress disorder.
- Digital Pills: Digital pills could help providers monitor patient compliance with medication in real time, potentially preventing serious outcomes.
- Digital Symptom Tracking: Symple's digital symptom tracking allows providers to identify patterns and alert patients in real time, helping to optimize mental health care for the future [6].

Conversational agents presented a promise to help with self-adherence and psychoeducation. Satisfaction rating was high across all studies [7]. Therapeutic mental health chatbots are AI agents that can be used as search assistants or recommendation systems, leading users to relevant mental health information or therapy content after a basic and brief dialogical interaction. Woebot, Wysa, and Tess are three of the most prominent chatbots that have emerged over the last few years, providing cognitive behavioral therapy in the form of brief, daily conversations, and mood tracking to help clients with depression and anxiety.

A randomized controlled trial study found that after two weeks of use, the Woebot group experienced a significant reduction in depression, as measured by the PHQ-9. Wysa employs several methods such as cognitive behavioral therapy, behavioral reinforcement, and mindfulness. Tess appears to have the most published research out of all these chatbot options [8]. After a quick scanning of Wysa and Woebot application, a comparison chart between Wysa and Woebot is shown in fig (9). Chezlong promises higher coherence and Arabic language conversation support with a lower price.

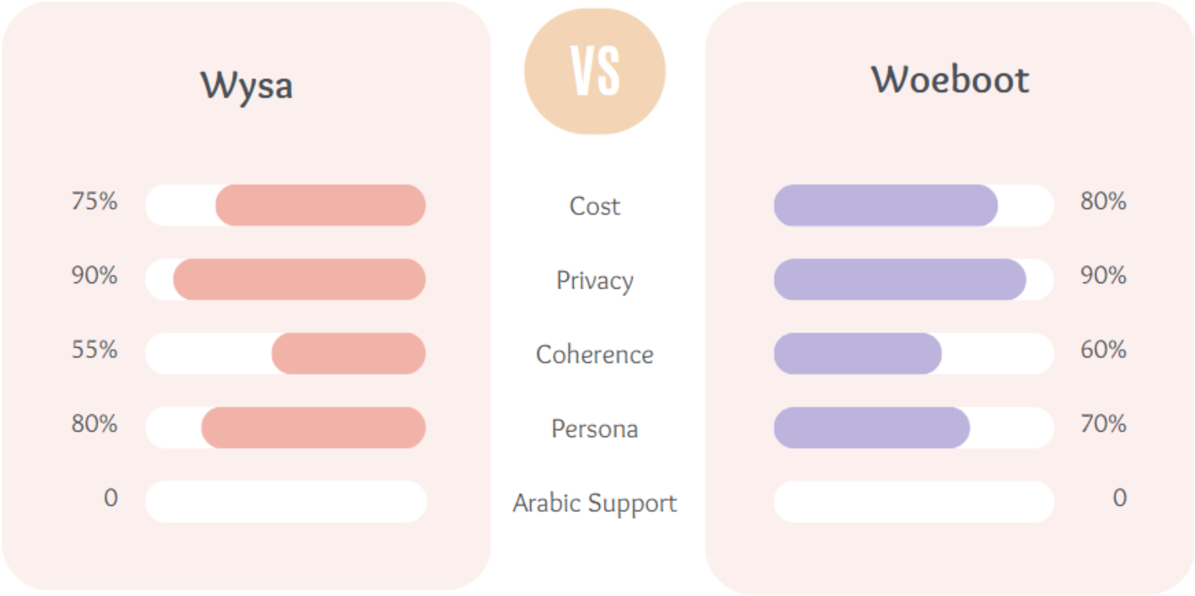


Figure (3): Comparison Chart Between Wysa and Woebot.

1.3. Objectives:

The project aims to develop an Arabic Chatbot dedicated to providing mental health support. The specific goals and objectives of this project revolve around addressing the growing need for accessible mental health resources in Arabic-speaking communities.

Firstly, the chatbot aims to offer a safe and confidential space for individuals to express their emotions and concerns without fear of judgment. Secondly, it seeks to provide accurate and reliable information about mental health conditions, treatment options, and coping strategies tailored to the cultural and linguistic nuances of Arabic-speaking users. Lastly, the chatbot aims to offer timely interventions and referrals to professional mental health services when necessary, ensuring that individuals receive the appropriate support and guidance they need.

In pursuit of these goals, the project intends to achieve several key outcomes. Firstly, it aims to develop a robust conversational AI system capable of understanding and responding to a wide range of user inquiries and emotional expressions in Arabic. This involves natural language processing techniques to comprehend user inputs and generate appropriate responses that are empathetic, informative, and culturally sensitive. Additionally, the project aims to integrate machine learning algorithms to continuously improve the Chatbot's responses based on user interactions and feedback. Furthermore, Chezlong will provide resources such as articles and exercises designed to promote mental well-being and resilience among Arabic-speaking users. Overall, the project seeks to empower individuals to take proactive steps towards managing their mental health effectively, thereby reducing stigma and increasing access to support services in Arabic-speaking communities.

By leveraging technology and linguistic expertise, the graduation project endeavors to make meaningful contributions to mental health care accessibility and awareness within Arabic-speaking populations. Through its user-centric design and personalized approach, Chezlong aims to foster a supportive environment where individuals feel understood, validated, and empowered to prioritize their mental well-being. Ultimately, the project aspires to serve as a valuable resource in destigmatizing mental health issues and promoting a culture of self-care and resilience in Arabic-speaking societies.

1.4. Scope:

The graduation project's boundaries and limitations revolve around the utilization of the Retrieval-Augmented Generation (RAG) model specifically for processing Arabic articles sourced from a predefined website. Included within the project's scope is the development and implementation of algorithms tailored for similarity search analysis and content generation in Arabic. This encompasses adapting RAG for Arabic language processing on a dataset of articles scraped from a designated website. The project does not include a user interface.

Conversely, features such as live chat support, personalized counseling, and real-time intervention are included in the project's limitations. By defining these boundaries, the project can focus on effectively utilizing RAG to provide valuable mental health support through curated Arabic content, aligning with its specified objectives and goals.

1.5. Significance of the Study:

The significance of the study lies in its potential to address a critical gap in mental health support within Arabic-speaking communities. The project offers a novel approach to enhancing accessibility and destigmatizing mental health issues. This endeavor is particularly significant given the cultural and linguistic barriers that often hinder individuals from seeking help or accessing relevant resources. By providing a confidential platform for individuals to express their emotions and access curated mental health content in their native language, the project aims to promote awareness, understanding, and acceptance of mental health challenges within Arabic-speaking societies.

Furthermore, the project's potential benefits extend beyond immediate support provision to broader applications in mental health care and research. The utilization of advanced natural language processing techniques, such as RAG, not only enables the Chatbot to understand and respond to emotional cues within Arabic text but also facilitates data collection and analysis on a large scale. This data-driven approach has the potential to yield valuable insights into prevalent mental health concerns, trends, and effective interventions within Arabic-speaking populations. Additionally, the Chatbot's ability to provide personalized recommendations based on users' emotional states can empower individuals to take proactive steps towards managing their mental well-being. Overall, the project's contributions have the potential to catalyze positive change in mental health awareness, accessibility, and support mechanisms for Arabic speakers, ultimately fostering a more inclusive and supportive environment for mental health care.

Chapter 2

Related Work

2.1 Introduction to Literature Review:

Systematically review and analyze existing research, studies, and resources related to mental health support, natural language processing techniques, and Arabic language processing. By conducting a comprehensive literature review, the project aims to gain insights into effective strategies for providing mental health support through chatbot technology, understanding the cultural and linguistic nuances of Arabic-speaking users, and identifying best practices for implementing natural language processing algorithms, such as the Retrieval-Augmented Generation (RAG) model. This review will serve as the foundation for developing a robust and culturally sensitive chatbot tailored to the needs of Arabic-speaking individuals seeking mental health support, thereby enhancing the effectiveness and accessibility of the intervention.

The chapter will delve into the first section on approaches for mental health support chatbots, where existing methodologies and studies on the effectiveness of chatbots in addressing mental health concerns will be reviewed, alongside discussions on design considerations. The subsequent section will focus on GPT models, introducing Generative Pre-trained Transformer (GPT) models and exploring their applications in natural language processing, particularly in mental health support and conversational agents. The chapter will then discuss the integration of approaches, synthesizing findings to propose a framework for developing effective mental health support chatbots using GPT models. Finally, the chapter will conclude by summarizing key insights, discussing implications for future research and practical applications, and emphasizing the significance of integrating approaches and leveraging GPT models in advancing mental health support through chatbot technology.

2.2 Approaches for Mental Health Support Chatbots:

Many studies have demonstrated a variety of methods and strategies for using NLP and ML in the case study of mental health. In this portion of the literature, I concentrated on case studies of chatbots.

a) Classical and Traditional Rule-Based Techniques:

One of the papers describes the process of implementing a chatbot for human mental health using a simple neural network. The goal is to enable people suffering from emotional disorders to improve their mental health in real time. The chatbot goes through several steps in the process of recognizing an emotional state, such as cleaning the data, identifying the emotional color in each word, and calculating the rating of each emotion. The output is presented in the form of a dictionary [9]. In a search for a different strategy, this article created a chatbot app to answer frequently asked questions/statements about mental health, using natural language processing, term frequency-inverse document frequency vectorization, principal component analysis, and ISOMAP algorithms. Jaccard similarity was chosen as the best algorithm for the chatbot due to its ability to handle questions and statements better than cosine similarity and Manhattan distance [10].

b) Deep Learning Techniques:

Inspired by the covid-19 pandemic, this paper discusses that chatbots can be a powerful tool for people facing mental health issues, providing a minimal loss and accuracy of 80.88% with Bidirectional LSTM model for sentiment analysis [11]. Another important and creative study proposes a deep learning-based chatbot framework to help mentally ill individuals identify their mental health conditions and provide appropriate therapy. It consists of three components: a keylogger module, a chat module, and a mental illness detection module. The keylogger collects data from the user's keyboard to track social media activity, while the chatbot converses with users and stores their daily chat history in real-time. The study also compared the accuracy of multiple deep learning classifiers such as Conv-LSTM and BERT for the Reddit Mental Health Dataset [12]. The objective of this last paper is to develop a chatbot techniques to promote their mental health through emotion recognition technique. Maxx, the chatbot, uses technologies like DialogFlow for Natural Language Processing (NLP), Flutter for app development and Google Cloud Platform (GCP) for data storage and security [13].

However, I focused my attention on this paper titled "Combating Depression in Students using an Intelligent ChatBot: A Cognitive Behavioral Therapy". Derived by the intuitive idea that social chatbots are those that develop a strong emotional connection with the user, the key concept of their work is creating an intelligent social therapeutic chatbot distributes text into emotion labels and identifies the users' mental state based on their chat data. It uses three deep learning classifiers, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Hierarchical Attention Network (HAN), to prevent pessimistic actions and rebuild constructive thoughts [14]. This paper aim to reduce mental illness in youth. The chatbot will ask questions to the user to understand the problem, identify emotions of the user to calculate the percentage of negativity in chat, and classify the level of mental status as normal, stressed, or depressed. To extract the emotion from the user chat data, the three well-known deep learning algorithms were deployed. This paper suggests a chatbot that can determine a user's mental state—such as normal, stressed, or depressed—by analyzing their dialogue. It has been trained to determine the percentage of each chat text that is positive and negative, as well as to categorize the text into emotions like Happy, Joy, Shame, Anger, Disgust, Sorrow, Guilt, and Fear. Fig (2) depicts the chatbot's proposed model.

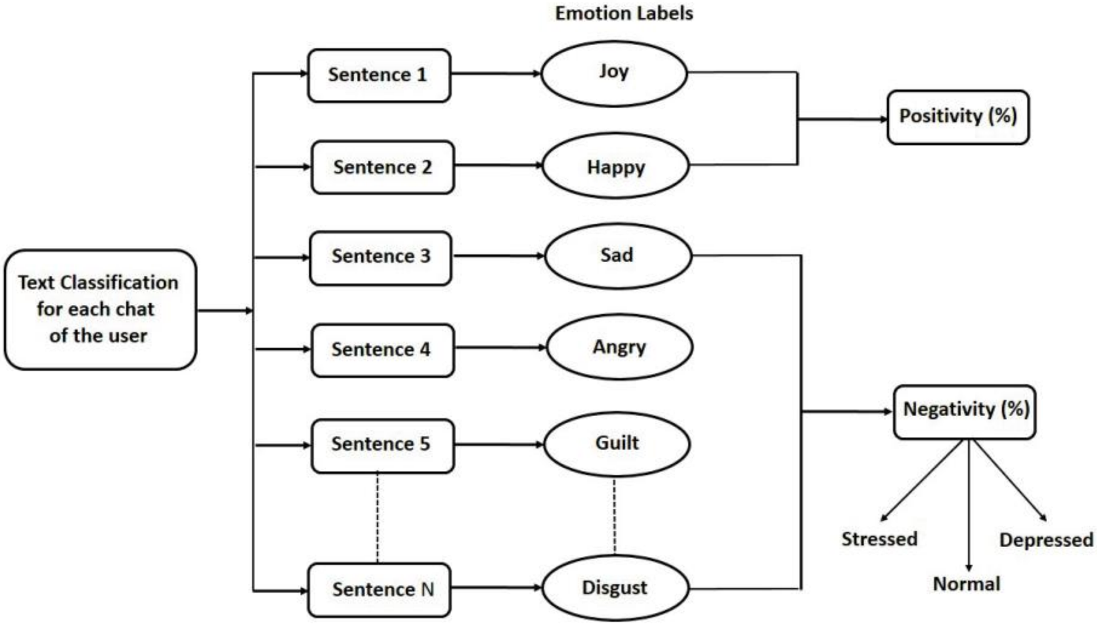


Figure (4): System model for mental state identification using chat data.

In order to identify emotions in text, they utilized the ISEAR dataset. The dataset includes 1542 emotional words and 7652 phrases. It can be divided into a number of broad emotional categories, including happiness, joy, shame, anger, disgust, sadness, guilt, and fear. A pipelined process for training and testing is shown in fig (3).

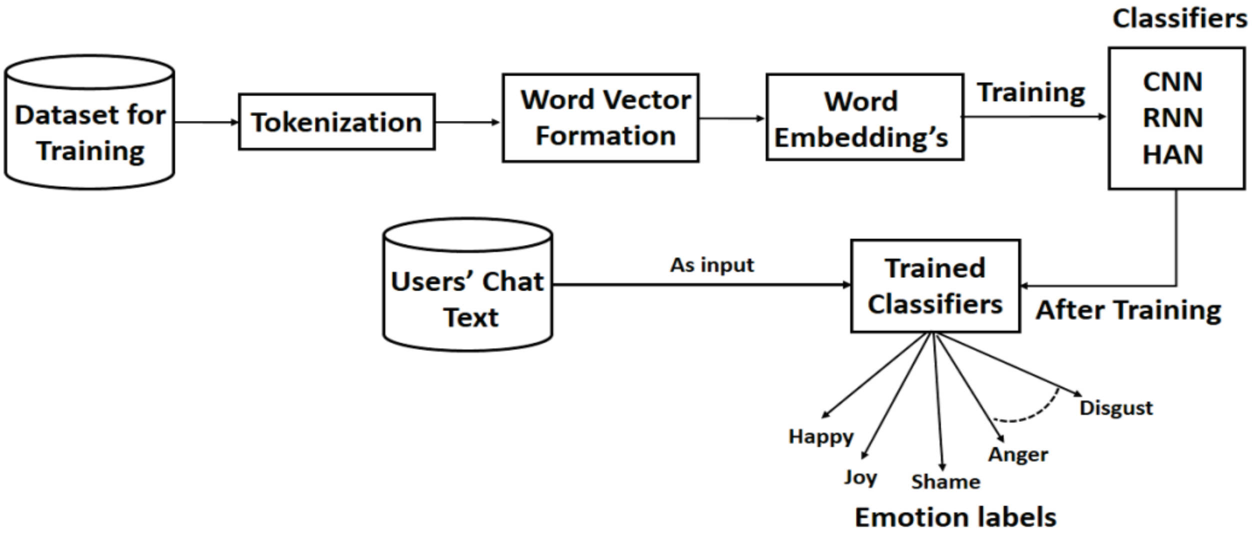


Figure (5): Pipelined process for training and testing to identify emotion label from users' chat.

a) Tokenization is the process of dividing a text into a list of sentences by using an unsupervised algorithm to build a model for abbreviations, words, collocations, and words that start sentences. The tool used is Punkt Sentence Tokenizer.

b) Forming Word Vector: Global Vectors for Word Representation (GloVe) is an unsupervised learning algorithm used to obtain vector representations for words. It is trained on aggregated global word-word cooccurrence statistics from a corpus, and the resulting represenations showcase interesting linear substructures of the wordvector space. The tools provided automate the collection and preparation of co-occurrence statistics for input into the model. The core training code is separated from these preprocessing steps and can be executed independently.

c) Word Embedding's: Neural network embedding is a mapping of a discrete categorical variable to a vector of continuous numbers. It is useful because it can reduce the dimensionality of categorical variables and meaningfully represent categories in the transformed space. It has three primary purposes: finding the nearest neighbor in space, forming cluster categories, and providing an input to machine learning models.

d) Classifiers: Three popular classifiers of deep neural networks are used for emotion classification from chat text: CNN, RNN, and HAN. CNN is a class of deep learning, feed-forward artificial neural networks that use multilayer perceptrons to detect patterns at every layer of convolution. RNN is a sequence of neural network blocks linked to each other like a chain. HAN is a Hierarchical Attention Network that captures two basic insights: document hierarchical structure and different informative words in a sentence. HAN has achieved high accuracy and performs better than CNN and RNN if we have a larger dataset.

Mental State Identification: Algorithm 1 takes user chat data as the input to identify the users' mental state. It uses Equation 1 and Equation 2 to calculate positivity and negativity percentage, which is then classified into five classes: normal, slightly stressed, highly stressed, slightly depressed, and highly depressed. If the negativity percentage is below 20, it is classified as normal, if it is in between 20-40, it is slightly stressed, and if it is between 40-60, it is highly stressed.

$$P(\%) = \frac{F(joy) + F(happy)}{Total\ no.of\ chat\ sentences\ (n)} \qquad (1)$$

$$N(\%) = \frac{F(sad)+F(angry)+..+F(disgust)}{Total\ no.of\ chat\ sentences\ (n)} \qquad (2)$$

CNN has achieved accuracy of up to 75% with high consistency for 15 epochs, while RNN and HAN have achieved up to 70% accuracy. CNN has outperformed the other two models in terms of training time, but HAN can perform better than CNN and RNN if we have a huge dataset. Algorithm 1 will classify the mental state of the user and advise them to take the mental treatment as follows: zero depression- No therapy requirement, mildly stressed- Relaxation required to shed stress, moderately stressed- Reduce stress in life, moderately depressed- Engage in recreational activities, and highly depressed- Meditation, relaxation is the need of the hour.

2.3 GPT Models:

The generative pre-trained transformer (GPT) is a family of language models trained on a large corpus of text data to generate human-like text. The transformer architecture's building blocks are used to construct them. Several natural language processing tasks, including text generation, language translation, and text classification, can be fine-tuned with it. Pre-training involves learning to anticipate the next word in a passage, which establishes a strong foundation for the model's performance on downstream tasks even in the absence of much task-specific data. ChatGPT, BioGPT, and ProtGPT2 have all used GPT language models [17]. A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighing the significance of each part of the input data. It is designed to process sequential input data, such as natural language. Transformers process the entire input all at once, allowing for more parallelization and therefore reducing training times. Transformers were introduced in 2017 by Google Brain and are increasingly the model of choice for NLP problems, replacing RNN models such as long short-term memory (LSTM). Pretrained systems such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) can be fine-tuned for specific tasks [18].

A. Attention is All You Need Review

Attention mechanisms have become an integral part of sequence modeling and transduction models, allowing modeling of dependencies regardless of their distance. This work proposes the Transformer, a model architecture neglecting recurrence and instead relying on an attention mechanism to draw global dependencies between input and output. It allows for significantly more parallelization and reached a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs. The Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence aligned RNNs or convolution. Self-attention is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. End-to-end memory networks are based on a recurrent attention mechanism instead of sequence-aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks [19]. The Transformer follows an encoder-decoder structure with stacked self-attention and pointwise, fully connected layers for both the encoder and decoder as in fig (4).

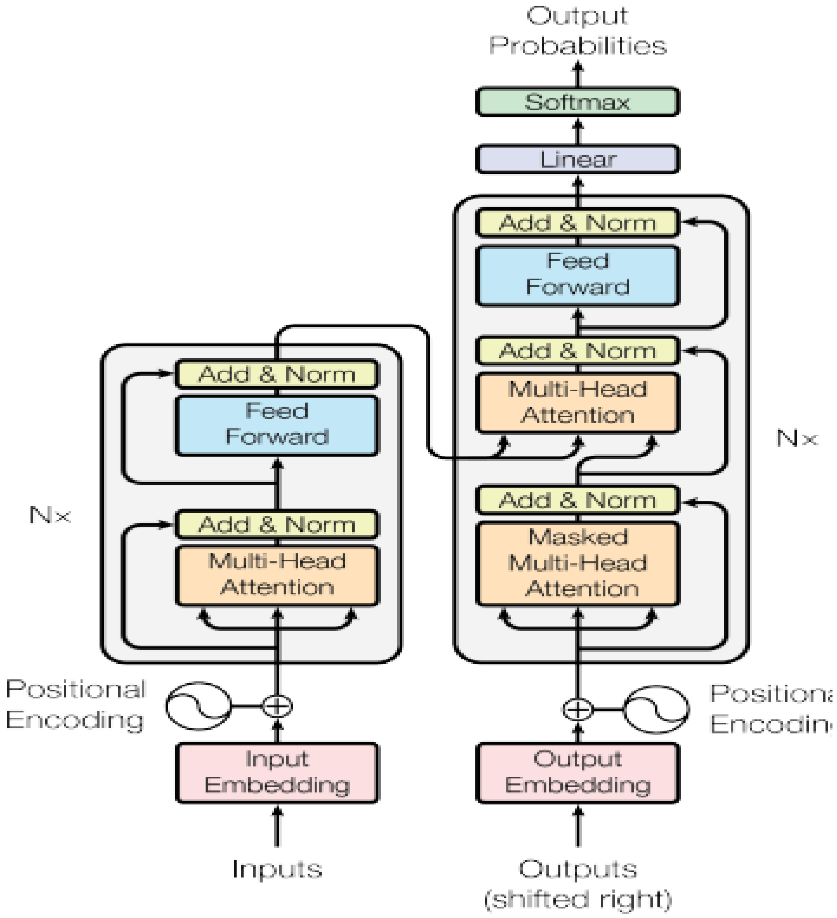


Figure (6): The Transformer - model architecture.

B. Improving Language Understanding by Generative Pre-Training Review

This paper explores a semi-supervised approach for language understanding tasks using a combination of unsupervised pre-training and supervised fine-tuning. They assume access to a large corpus of unlabeled text and several datasets with manually annotated training examples (target tasks). They employ a two-stage training procedure, using a language modeling objective on the unlabeled data to learn the initial parameters of a neural network model. For model architecture, they use the Transformer.

During transfer, they utilize task-specific input adaptations derived from traversal-style approaches. Our general task-agnostic model outperforms discriminatively trained models that employ architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. They achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI). The training process is divided into two steps. A high-capacity language model is first learned using a large text corpus. The model is then tuned in a subsequent stage using labelled data to perform a discriminative task [20] fig (5).

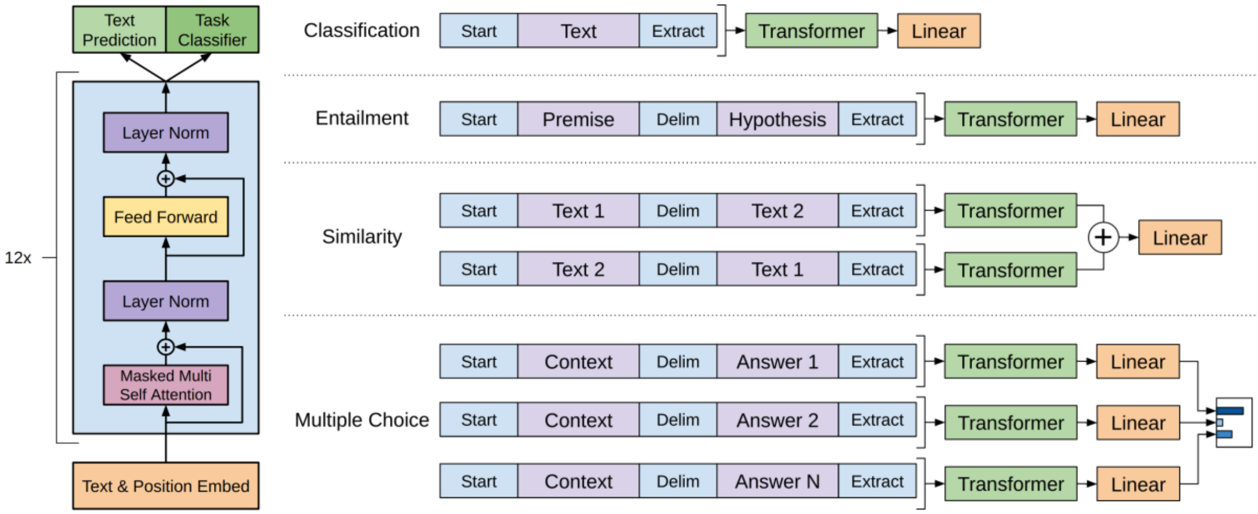


Figure (7): (left) Transformer architecture and training objectives used in this work. (right) Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Chapter 3

Materials and Methods

3.1 System Description:

This system diagram illustrates the basic components and workflow of a chatbot working with RAG for mental health support, demonstrating how it processes user inputs, retrieves relevant information, analyzes sentiment, generates responses, and delivers support to users fig(8).

1) Chatbot Interface: This is the interface through which users interact with the chatbot. It can be a website or a mobile app.
2) User Input Processing: When a user sends a message or query, the input is processed by the chatbot. This step involves natural language processing techniques to understand the user's intent and extract relevant information from the text.
3) Retrieval System: The chatbot uses a retrieval system to search through a database or corpus of pre-existing articles for most related information to the user’s query. These articles serve as a knowledge base from which the Large Language Model (LLM) can draw responses.
4) Prompt Augmentation: User’s query is compounded with the source knowledge to design prompt. This enables the LLM to generate a response using the RAG (Retrieval-Augmented Generation). This model combines elements of both retrieval-based and generative-based approaches to produce contextually relevant and coherent responses fig (8).
5) Response Delivery: The generated response is then delivered back to the user through the chatbot interface. This response may include information, advice, or resources related to the user's query and emotional state, providing support and guidance in managing their mental health.

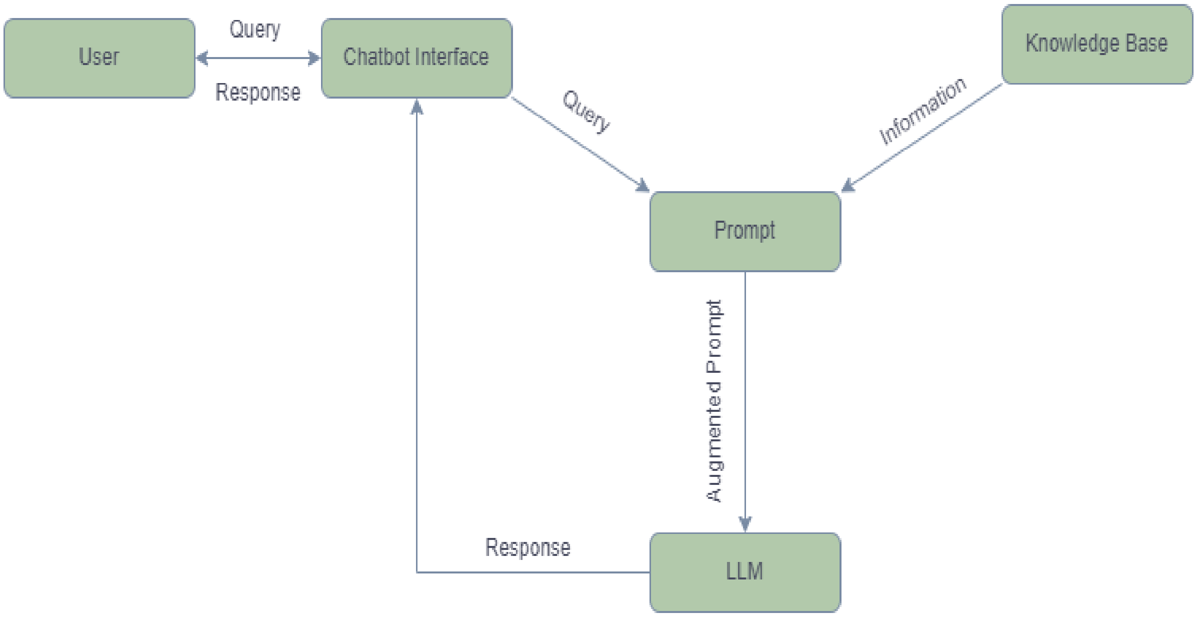
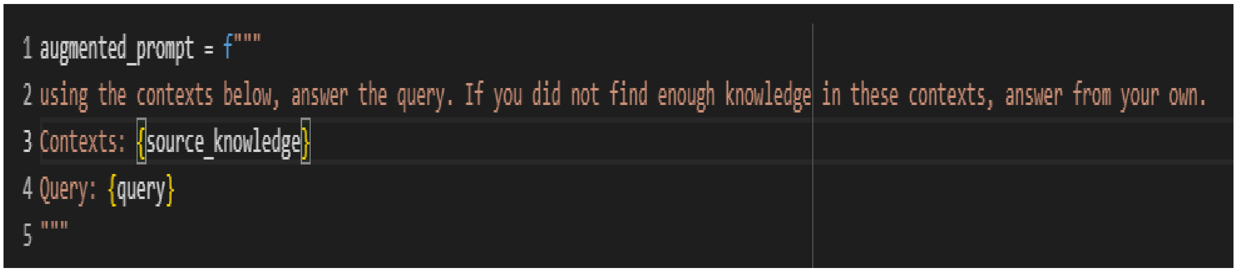


Figure (8): System Design – Abstract View.



```
augmented_prompt = f"""
using the contexts below, answer the query. If you did not find enough knowledge in these contexts, answer from your own.
Contexts: {source_knowledge}
Query: {query}
"""
```

Figure (9): Augmented Prompt.

3.2 System Requirements:

The system requirements for the chatbot encompass hardware, software, and data considerations necessary for its development, deployment, and operation. Firstly, from a hardware perspective, the chatbot requires computing resources like GPU to execute its algorithms efficiently. This includes a server or cloud infrastructure capable of handling natural language processing tasks with sufficient processing power and memory to manage user interactions in real-time. Additionally, the hardware should support scalability to accommodate potential increases in user traffic and data processing demands over time. Furthermore, the chatbot may integrate with web services like OpenAI, Hugging Face, and Pinecone to apply various tasks through the pipeline of the generative process. Meeting these system requirements ensures the chatbot's reliability, scalability, and effectiveness in delivering mental health support to Arabic-speaking users.

3.3 Design Constraints:

Design constraints for the chatbot encompass various factors that influence its development, functionality, and usability. Firstly, linguistic constraints pose challenges due to the complexity of the Arabic language, including dialectal variations, morphology, and syntax. Designing the chatbot to understand and generate coherent responses in different Arabic dialects while maintaining cultural sensitivity requires specialized linguistic expertise and robust natural language processing algorithms tailored to Arabic language processing. Technical constraints include limitations in computational resources, such as processing power, memory, and bandwidth, which may restrict the chatbot's functionality and scalability. Ethical and regulatory constraints also pose challenges, particularly concerning user privacy, data security, and compliance with healthcare regulations.

3.4 Detailed Design:

The detailed design of the RAG (Retrieval-Augmented Generation) model incorporates several components optimized for efficient similarity search and conversational schema generation. The design leverages the paraphrase-multilingual-MiniLM-L12 embedding model as the foundation for similarity analysis, enabling the generation of embeddings that capture semantic similarities between user’s query and the knowledge base.

The data is processed using Pinecone, a vector storage service optimized for similarity search tasks. Pinecone allows for the storage and retrieval of high-dimensional embeddings efficiently, making it suitable for handling large-scale datasets with high-dimensional embeddings. In this configuration, the embeddings are stored in Pinecone with a dimensionality of 384 and cosine distance metric, enabling fast and accurate similarity search operations [30].

Additionally, the RAG model incorporates the GPT-3.5-turbo language model (LLM) for conversational schema generation. GPT-3.5-turbo is a variant of the GPT-3.5 model, enhanced for improved performance and efficiency. The language model is passed to ChatOpenAI, a conversational AI service, to generate conversational responses based on the provided schema. The temperature parameter is set to 0.5, controlling the randomness and creativity of the generated responses [31]. Before ChatOpenAI step, the messages are passed to a list containing all the historical messages. There are three types of messages: SystemMessage, HumanMessage, and AIMessage. This also works as a memory for the chatbot. The SystemMessage is written as follows:

'أنت معالج نفسي. اسمك تشيزلونج ويمكن أن ندعوك تشيز فقط. ستساعدني في تخطي الأوقات الصعبة
وتزويدي بمعلومات صحية. ابدأ بقول "السلام عليكم" ثم تعريف نفسك.'

Overall, the detailed design of the RAG model combines state-of-the-art embedding models, vector storage technology, and language models to enable efficient similarity search analysis and conversational schema generation. By leveraging these components, the RAG model offers enhanced capabilities for a wide range of natural language processing tasks, including information retrieval, question answering, and dialogue generation.

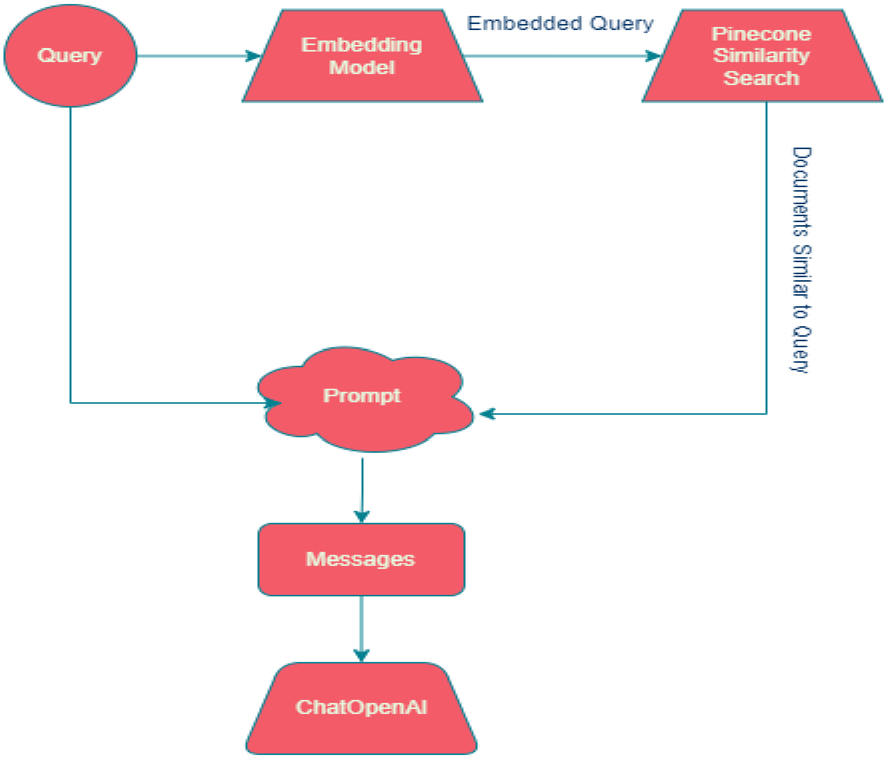


Figure (10): RAG Chat Design.

3.5 Data Design:

"nafsy" arabic dataset is used as a knowledge base for this project. This dataset is available on Kaggle [21]. It was originally scrapped from nafsy.net website. It is a comprehensive collection of Arabic text data related to various psychological topics. It consists of 664 sample articles and blog posts written in Arabic covering subjects such as mental health, well-being, therapy, and personal development. The dataset is curated to provide valuable insights into the psychological landscape within Arabic-speaking communities.

3.5.1 Original Dataset:

The dataset exhibits a diverse distribution of text sizes, ranging from a minimum of 398 characters to a maximum of 38,696 characters. On average, the articles have a length of approximately 5,967.48 characters, with a standard deviation of 4,739.43 characters. This variation in text sizes reflects the diverse nature of the content included in the dataset, ranging from short blog posts to more in-depth articles discussing complex psychological concepts fig (11).

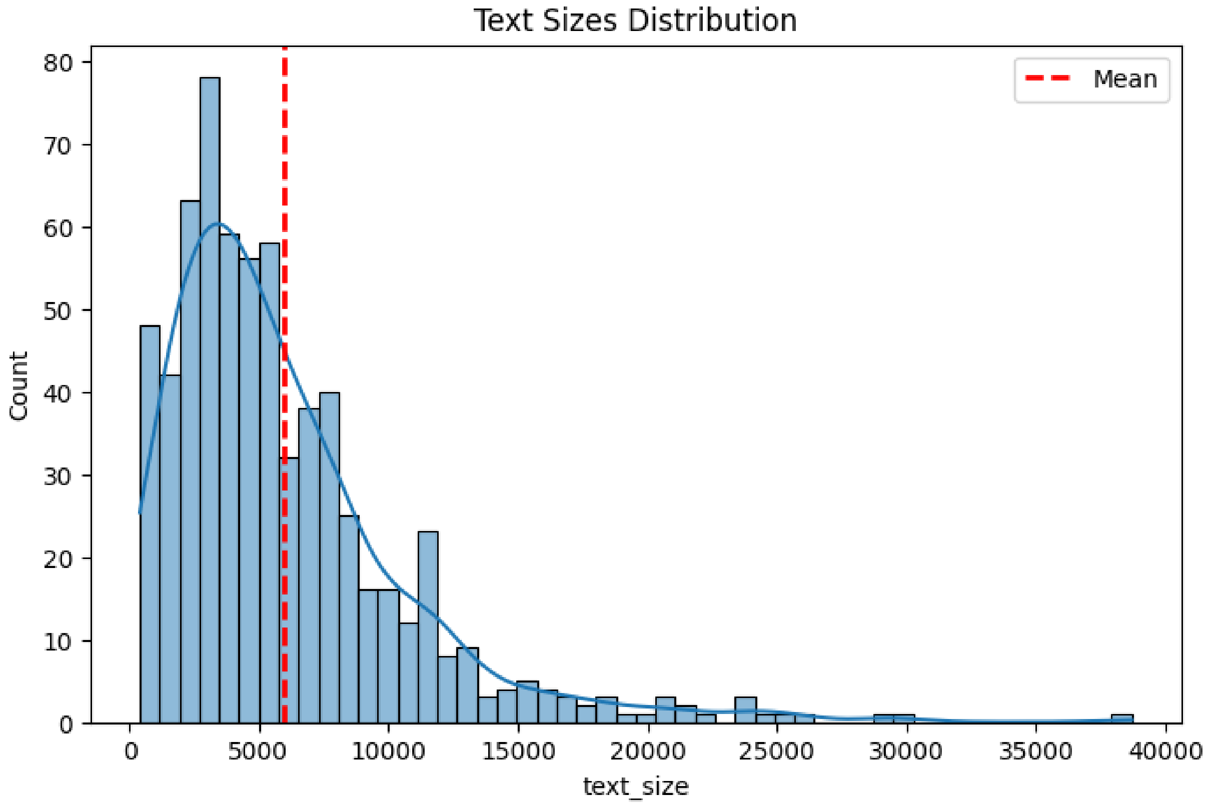


Figure (11): Histogram of text sizes of Nafsy dataset.

3.5.2 Splitting Into Smaller Chunks:

Moreover, the dataset has been split into smaller chunks. Splitting the dataset into smaller chunks is a necessary step for several reasons:

1) Efficient Processing and Parallelization: Large datasets can be computationally intensive to process, especially when conducting analyses or training machine learning models. By splitting the dataset into smaller chunks, researchers can distribute processing tasks across multiple machines or processors, leading to faster execution times and improved efficiency. This parallelization can significantly accelerate tasks such as data preprocessing, feature extraction, or model training, enhancing productivity and scalability.
2) Memory Constraints: Some machine learning algorithms or analysis techniques may require loading the entire dataset into memory, which can be challenging for large datasets. Splitting the dataset into smaller chunks reduces memory requirements, allowing researchers to work with subsets of the data without encountering memory issues.
3) Model Validation: When developing machine learning models, it's essential to partition the dataset into training, validation, and test sets. Splitting the dataset into smaller chunks enables researchers to create multiple partitions with sufficient data samples for training, validation, and testing, ensuring robust model evaluation and generalization performance.
4) Data Exploration and Analysis: Splitting the dataset into smaller chunks allows researchers to explore and analyze different subsets of the data systematically. This approach enables targeted analyses on specific subsets or segments of the dataset, facilitating deeper insights into particular themes, trends, or patterns within the data.

Chunking has been done using Langchain TokenTextSplitter function with parameters: chunk_size = 2048 and chunk_overlap = 20. LangChain is a framework for building language model-powered applications that are context-aware and capable of reasoning. It integrates language models with contextual sources and enables applications to make informed decisions based on provided context [22]. The TokenTextSplitter divides a raw text string by initially transforming the text into Byte Pair Encoding (BPE) tokens. Subsequently, it segments these tokens into segments and reverts the tokens within each segment back into text [23].

The chunking process resulted in the creation of a dataset comprising 2,260 chunks of text. These chunks exhibit a distribution of text sizes, with a minimum chunk size of 22 characters, a maximum size of 7,164 characters, and a mean size of approximately 1,767.49 characters. Additionally, the dataset has a standard deviation of 574.26 characters, indicating a moderate level of variability in chunk sizes.

Notably, the dataset was shuffled, meaning that the order of chunks was randomized to prevent any inherent biases or patterns in the original sequence from affecting subsequent analyses or processing. This shuffled dataset of text chunks provides a structured and varied corpus for further exploration, analysis, or training of language models and other natural language processing tasks fig (12).

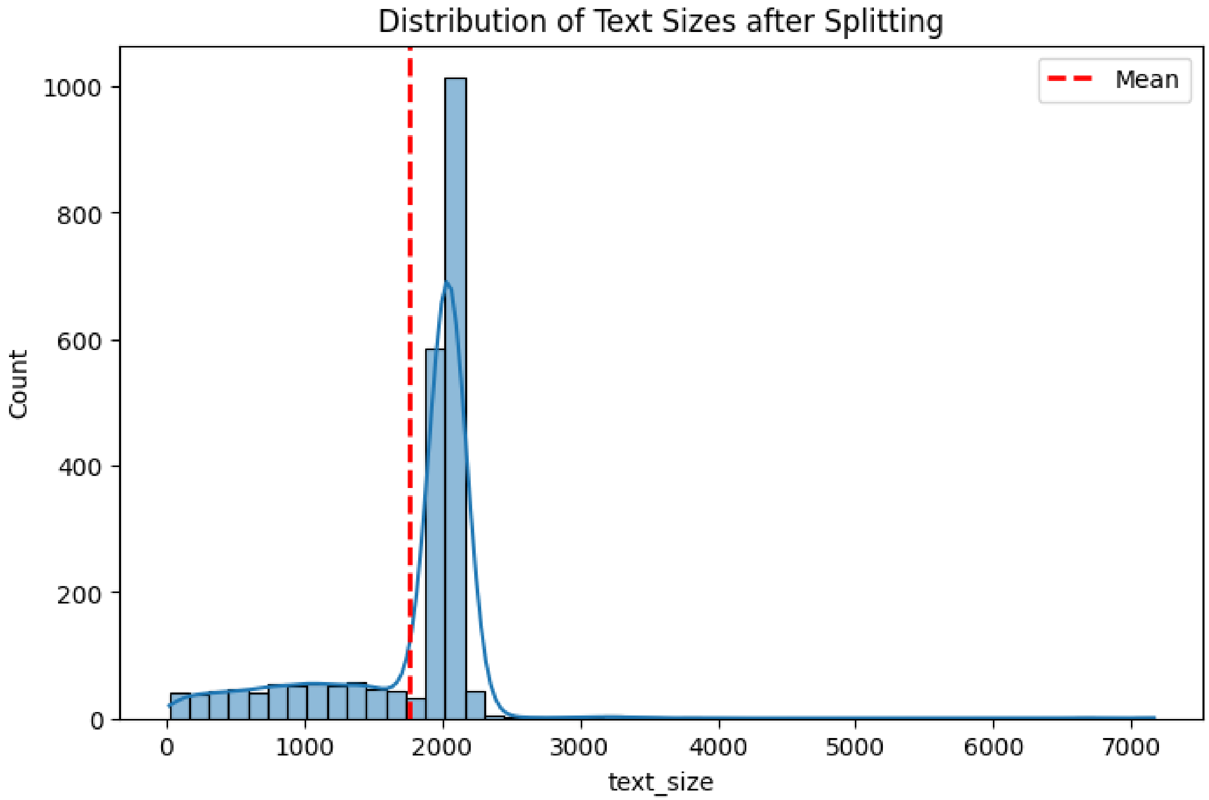


Figure (12): Histogram of text sizes of Chunked Nafsy dataset.

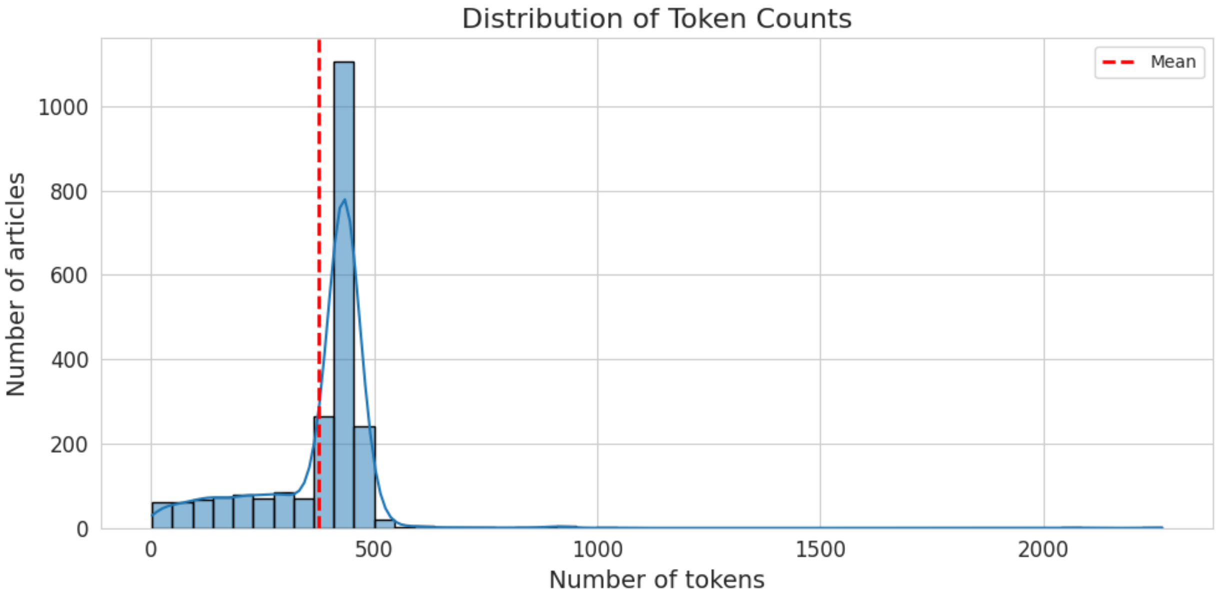


Figure (13): Tokens Distribution of Nafsy Datastet.

3.5.3. Cleaning:

The next step in the process involved cleaning the text data. The cleaning process involved several steps: removing URLs, detaching punctuations, eliminating extra spaces, normalization, and dropping duplicates (resulting in 2,216 chunks). This cleaning is vital for preparing the Arabic text data for further analysis and processing. Detaching punctuation ensures that the text is parsed correctly and accurately, preventing punctuation marks from interfering with subsequent processing steps. Eliminating extra spaces enhances the readability and consistency of the text, making it easier to interpret and analyze. Additionally, dropping duplicates helps reduce redundancy and ensures that each text sample in the dataset is unique, thereby preventing duplication of information and optimizing resource utilization. Overall, these cleaning steps contribute to enhancing the quality, consistency, and usability of the Arabic text data, enabling more effective analysis and modeling in natural language processing tasks.

In this process, I utilized Python Natural Language Tool Kit (NLTK) library and regular expression library (re in python). NLTK is designed for natural language processing tasks. It offers a comprehensive set of tools and resources for tasks such as tokenization, stemming, part-of-speech tagging, and parsing [24]. The re library in Python provides support for working with regular expressions, which are powerful tools for pattern matching and string manipulation. With re, developers can perform various tasks such as searching, extracting, and replacing specific patterns within text data [25].

3.5.4 Exploratory Text Analysis with BERTopic:

Further Exploratory Text Analysis is done using BERTopic. BERTopic is an unsupervised topic modeling technique that utilizes BERT embeddings, a pre-trained language model, to extract topics from a corpus of text documents. Unlike traditional methods, BERTopic captures semantic relationships between words and documents through contextual embeddings [26]. The benefit of this step lies on giving:

1) Understanding Document Collections: Topic modeling allows for the exploration and understanding of Nafsy collections by uncovering the underlying themes and topics present in the data.
2) Content Organization: Topic modeling helps in organizing large document collections by grouping similar documents together based on their thematic content. This facilitates efficient retrieval and browsing of documents, making it easier to navigate through large corpora.
3) Information Retrieval: Topic modeling enhances information retrieval by providing a structured representation of document content. Users can search for documents based on specific topics or themes of interest, improving the relevance and accuracy of search results.
4) Topics can be represented as metadata for the retrieved information.
5) Identifying Trends and Patterns: Topic modeling helps in identifying trends and patterns in Nafsy dataset by uncovering recurring themes and topics.

For BERTopic modeling, it is necessary to transform the text into numerical representations, that is embeddings. Text embeddings are numerical representations of words, sentences, or documents in a high-dimensional vector space. These embeddings capture semantic relationships between words or texts, allowing for meaningful comparisons and computations. In the context of natural language processing (NLP), text embeddings are often generated using neural network models trained on large text corpora. These models learn to encode semantic information about words or texts into dense vector representations, where similar words or texts are represented by vectors that are closer together in the embedding space.

The embedding model used here is asafaya/bert-base-arabic [27], which is a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model specifically trained on Arabic text data. By utilizing the asafaya/bert-base-arabic model, BERTopic is able to generate embeddings for Arabic text data, enabling the extraction of meaningful topics from Arabic documents. The embeddings generated by the BERT model are stored using the pickle format. Pickle is a Python module that serializes Python objects into a binary format, allowing for efficient storage and retrieval of complex data structures such as embeddings. It enables the embeddings generated by BERTopic to be saved to disk and loaded back into memory when needed, facilitating fast and seamless processing of text data in topic modeling tasks.

The pipeline stack of the BERTopic constists of fig(14):

1) Uniform Manifold Approximation and Projection (UMAP): UMAP is a nonlinear dimensionality reduction technique that is commonly used to visualize high-dimensional data in lower-dimensional spaces while preserving local and global structure. The following hyperparameters are specified:

    - (n_neighbors) to control the number of neighboring points used to construct the local neighborhood for each data point. In this case, n_neighbors is set to 7, meaning that each point will be connected to its 7 nearest neighbors in the high-dimensional space.

    - (n_components) to determine the number of dimensions in the reduced space. A lower value for n_components results in a more compressed representation of the data. Here, n_components is set to 45, indicating that the dimensionality of the data will be reduced to 45 dimensions.

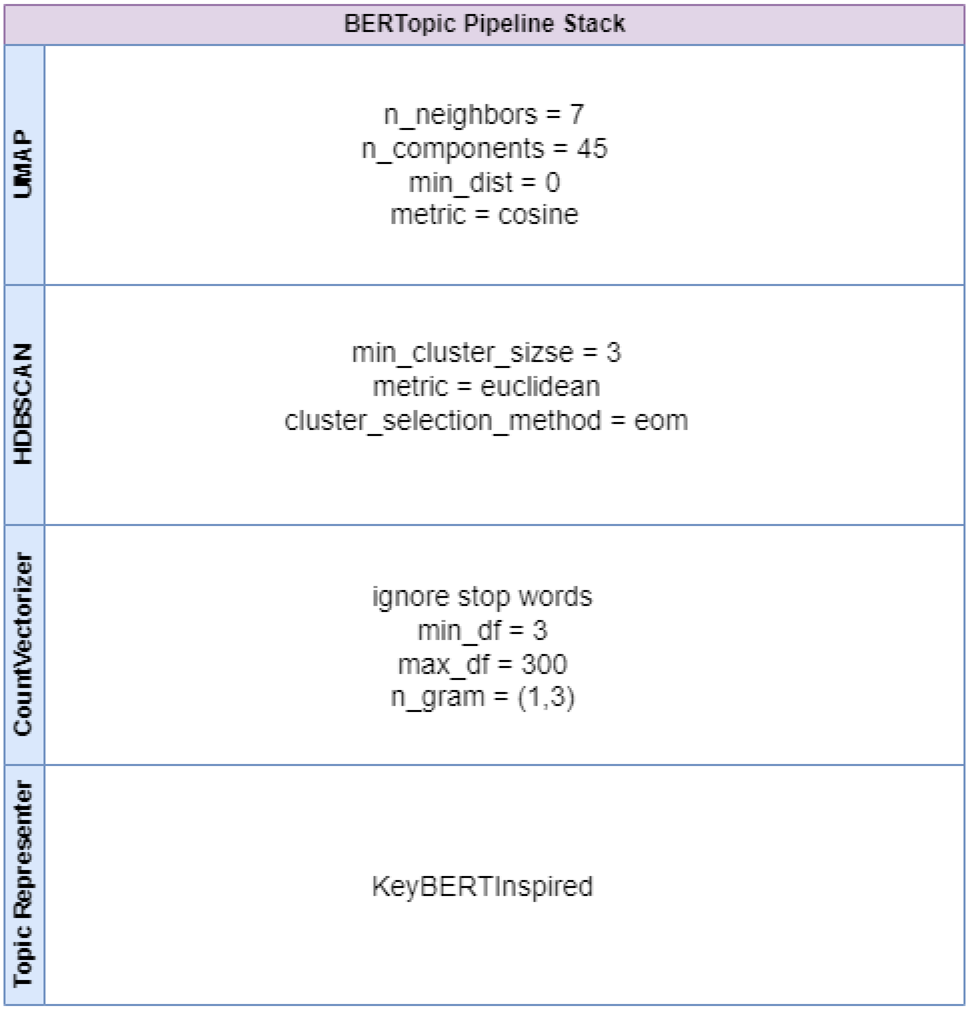
- (min_dist) to control the minimum distance between points in the reduced space. A lower value for min_dist encourages tighter clustering of points in the reduced space. In this case, min_dist is set to 0, indicating that there is no enforced minimum distance between points.
- (metric) to specifie the distance metric used to measure the similarity between data points. In this case, the cosine distance metric is used, which measures the cosine of the angle between two vectors. This metric is well-suited for text data as it captures the semantic similarity between documents.

2) HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise): HDBSCAN is a density-based clustering algorithm that is capable of discovering clusters of varying shapes and densities in high-dimensional spaces. By applying HDBSCAN, the BERTopic pipeline aims to partition the data into clusters based on their density and similarity, ultimately facilitating the extraction of coherent and meaningful topics from the text data. The following hyperparameters are specified:
    - min_cluster_size: This parameter determines the minimum number of samples required to form a cluster. Clusters with fewer samples than min_cluster_size are treated as noise points. Here, min_cluster_size is set to 3, meaning that clusters must contain at least 3 samples to be considered valid.
    - metric: This parameter specifies the distance metric used to measure the similarity between data points. In this case, the Euclidean distance metric is used, which measures the straight-line distance between two points in Euclidean space.
    - cluster_selection_method: This parameter determines the method used to select clusters from the hierarchy generated by HDBSCAN. The 'eom' (Excess of Mass) method is used here, which selects clusters based on their density and separation from noise points. This method tends to produce more compact clusters with well-defined boundaries.

3) CountVectorizer: CountVectorizer is a feature extraction technique commonly used in natural language processing to convert text data into numerical vectors suitable for machine learning algorithms. It represents each document as a vector where each element corresponds to the frequency of a particular word or n-gram in the document. In this step, CountVectorizer is imported from the sklearn.feature_extraction module. The ignore stop words parameter is applied, which means that common words such as "و," "من," and "أن" are excluded from the vocabulary as they typically do not carry significant meaning for topic modeling. Additionally, the following hyperparameters are specified:

   - min_df: This parameter specifies the minimum frequency threshold for a word to be included in the vocabulary. Words that occur less frequently than min_df across all documents are ignored. Here, min_df is set to 3 as these words might be noise.
   - max_df: This parameter specifies the maximum frequency threshold for a word to be included in the vocabulary. Words that occur more frequently than max_df across all documents are ignored. Here, max_df is set to 300 and treated as stop words.
   - ngram_range: This parameter specifies the range of n-grams to be considered during vectorization. An n-gram is a contiguous sequence of n items (words or characters) from a given sample of text. In this case, ngram_range is set to (1, 3), meaning that unigrams (single words), bigrams (pairs of consecutive words), and trigrams (sequences of three consecutive words) are considered during vectorization.

4) Topic Representer: which is inspired by KeyBERT. KeyBERT is a keyword extraction model based on BERT embeddings that identifies the most representative keywords or key phrases for a given document or topic.



BERTopic Pipeline Stack	

UMAP	n_neighbors = 7
	n_components = 45
	min_dist = 0
	metric = cosine

HDBSCAN	min_cluster_sizse = 3
	metric = euclidean
	cluster_selection_method = eom

CountVectorizer	ignore stop words
	min_df = 3
	max_df = 300
	n_gram = (1,3)

Topic Representer	KeyBERTInspired

Figure (14): BERTopic Pipeline Stack.

3.5.5 Data Deduplication:

A deeper level of data preprocessing was Data Deduplication. Data deduplication is a process in which duplicate or redundant entries within a dataset are identified and removed, resulting in a dataset with unique and non-repetitive samples. The redundancy and duplication are decided by using certain semantic similarity algorithm and controlled by a certain threshold.

Initially, the data was tokenized using the asafaya/bert-base-arabic tokenizer, which is specifically trained for Arabic text. Once tokenized, the data was embedded using the sentence-transformer/paraphrase-multilingual-MiniLM-L12-v2 embedding model. The sentence-transformer/paraphrase-multilingual-MiniLM-L12-v2 model is a pre-trained transformer-based model designed for multilingual text understanding tasks. It is trained to generate embeddings that capture semantic similarities between sentences in multiple languages, making it suitable for applications involving multilingual text data [28].

After embedding, the data was processed using Faiss, an efficient similarity search library, for embedding storage. Faiss enables fast similarity search and clustering of high-dimensional embeddings, making it ideal for handling large-scale datasets with high-dimensional embeddings efficiently.

Finally, the data deduplication step was applied to remove duplicate entries from the dataset. This ensures that each sample in the dataset is unique, preventing redundancy and improving the quality of the data for subsequent analysis or modeling tasks. After deduplication, the final size of the dataset was reduced to 1884 samples, each representing a unique and non-repetitive entry [29].

Chapter 4

Implementation and Preliminary Results

4.1 Programming Languages and Tools:

The chatbot implementation utilizes the following programming languages, frameworks, and tools:

1) Python: Python serves as the primary programming language for the chatbot implementation. It offers a rich ecosystem of libraries and tools for natural language processing (NLP), machine learning, and web development, making it well-suited for building sophisticated chatbot systems.

2) Langchain Library: Langchain is a framework used for developing applications powered by language models. It facilitates building context-aware applications that can reason and generate responses based on provided context. Langchain enables seamless integration with language models and provides an intuitive interface for developing conversational AI systems.

3) re Library: The re library in Python is used for regular expression operations. It provides functionalities for pattern matching and string manipulation, which are essential for text preprocessing and data cleaning tasks in the chatbot implementation.

4) NLTK Library: NLTK (Natural Language Toolkit) is a popular library for natural language processing in Python. It offers a wide range of tools and resources for tasks such as tokenization, stemming, part-of-speech tagging, and parsing. NLTK is chosen for its extensive capabilities and ease of use in processing and analyzing text data.

5) Seaborn and Matplotlib Libraries: Seaborn and Matplotlib are Python libraries used for data visualization. They provide powerful tools for creating various types of plots and graphs, allowing for effective visualization of chatbot performance metrics, user interactions, and other relevant data.

6) OpenAI API: OpenAI provides an API for accessing state-of-the-art language models such as GPT (Generative Pre-trained Transformer) models. The OpenAI API is utilized for tasks such as conversational response generation and language understanding, leveraging the advanced capabilities of these language models.
7) Hugging Face Models: Hugging Face offers a repository of pre-trained transformer-based models for natural language understanding and generation tasks. These models are extensively used in the chatbot implementation for tasks such as text generation, text embedding, similarity search, and topic modeling.
8) Pinecone and Faiss for Vector Storage: Pinecone and Faiss are vector storage services optimized for similarity search tasks. They offer efficient storage and retrieval of high-dimensional embeddings, making them ideal for handling large-scale datasets and enabling fast and accurate similarity search operations in the chatbot implementation.
9) Sentence Transformers and BERTopic Modeling: Sentence Transformers and BERTopic are libraries for generating embeddings and performing topic modeling tasks, respectively. They are utilized in the chatbot implementation for tasks such as text representation learning, semantic similarity analysis, and topic extraction from text data.

The choice of these technologies is justified by their robustness, scalability, and suitability for building advanced conversational AI systems. They offer a comprehensive set of features and functionalities required for processing, analyzing, and generating natural language text, enabling the development of a powerful and effective chatbot solution.

4.2 BERTopic Modeling Results:

The results of the BERTopic modeling reveal several key insights into the distribution of topic predictions across the dataset. With a total of 187 topics identified, the probability distribution of predictions provides valuable information about the confidence levels associated with each topic assignment fig (15).

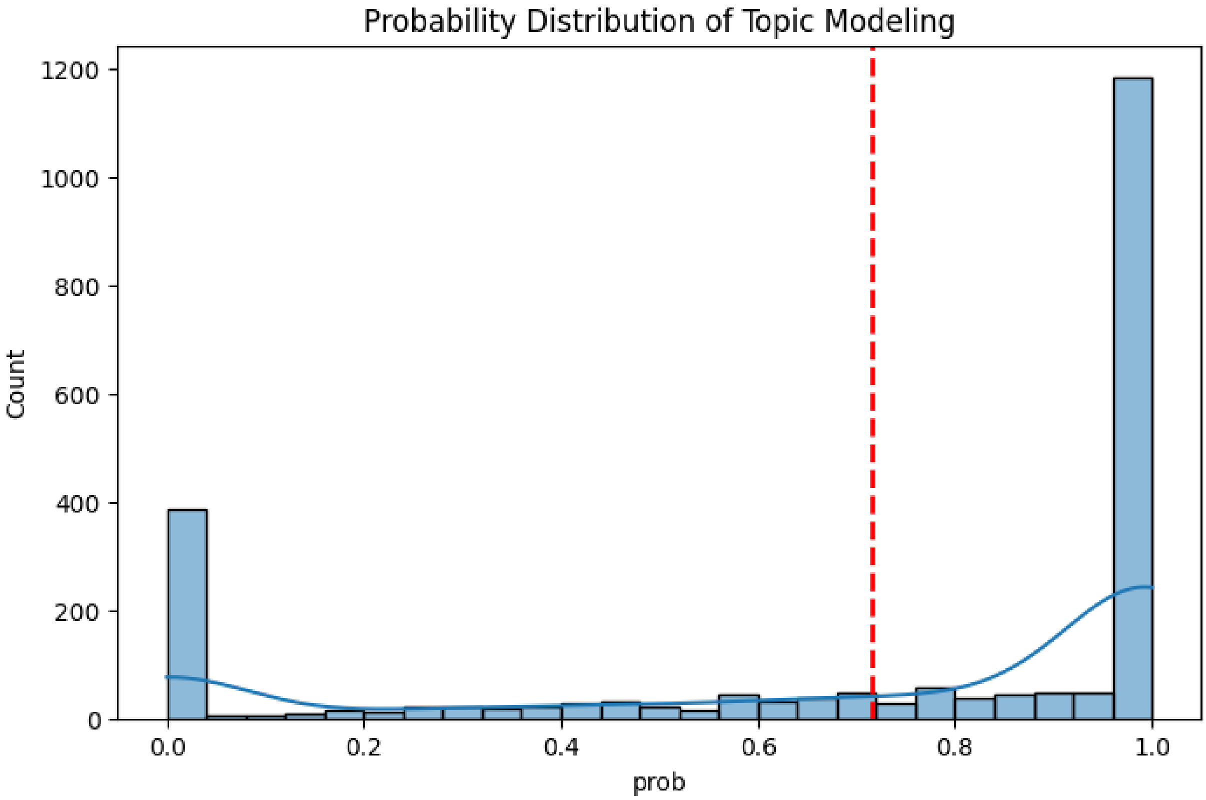


Figure (15): Probability distribution of Topic Modeling.

The mean probability of topic prediction is calculated to be approximately 0.716, indicating a relatively high average confidence level in the assigned topics. This suggests that, on average, the BERTopic model is confident in its ability to accurately assign topics to the input data.

However, the standard deviation of the probability distribution is calculated to be approximately 0.387, indicating a considerable degree of variability in the confidence levels across different topic predictions. This variability suggests that while some topics are confidently assigned with high probabilities, others may have lower confidence levels or be more uncertain.

The quartile analysis further elucidates the distribution of probability values. The 25th quartile (Q1) is calculated to be approximately 0.456, indicating that at least 25% of the topic predictions have probabilities below this threshold. Similarly, the 75th quartile (Q3) is calculated to be 1, indicating that at least 75% of the topic predictions have probabilities equal to or below this value. The 50th quartile (Q2), also known as the median, is calculated to be 1, indicating that 50% of the topic predictions have probabilities equal to or below this value.This suggests that a significant portion of the topic predictions are associated with high confidence levels, as indicated by a probability of 1.

Overall, the probability distribution of topic predictions provides insights into the confidence levels associated with each topic assignment, highlighting both the overall average confidence level and the variability in confidence levels across different topics. These insights can inform further analysis and interpretation of the topic modeling results, guiding decisions on the relevance and significance of individual topics within the dataset.

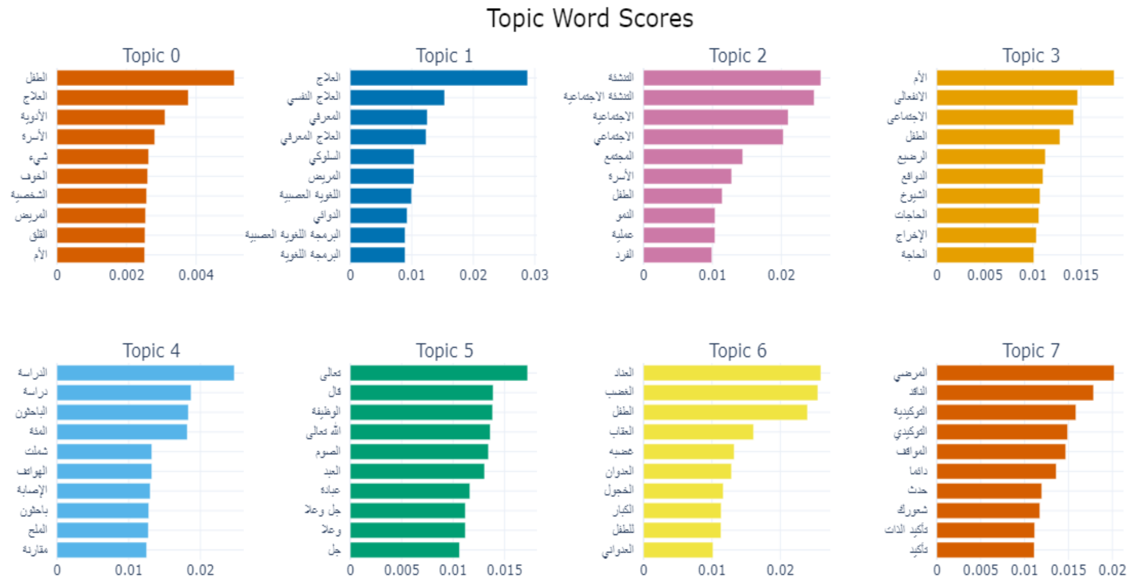


Figure (16): Topic Word Scores.

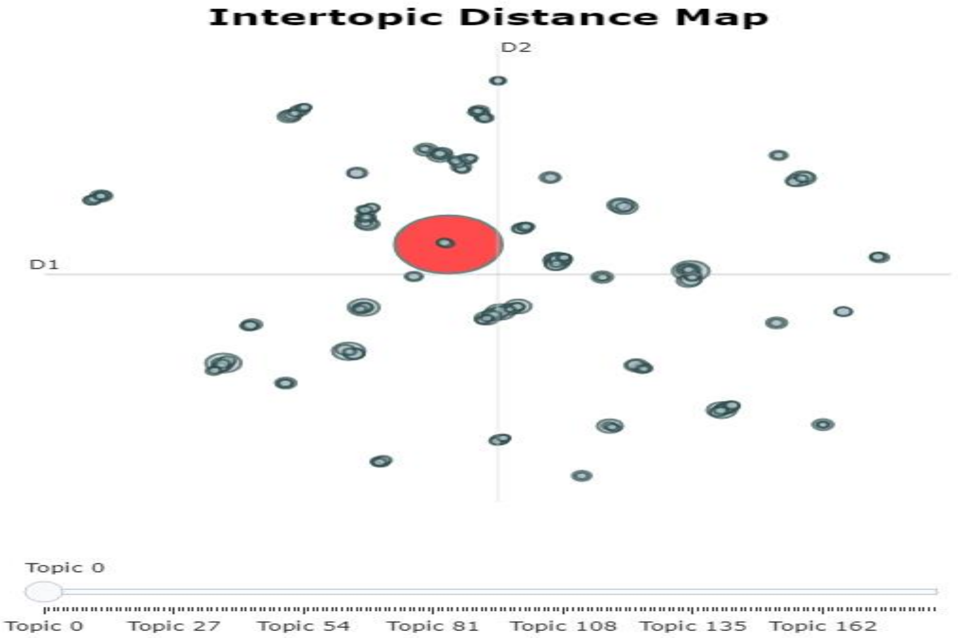


Figure (17): Intertopic Distance Map of Topic Modeling.

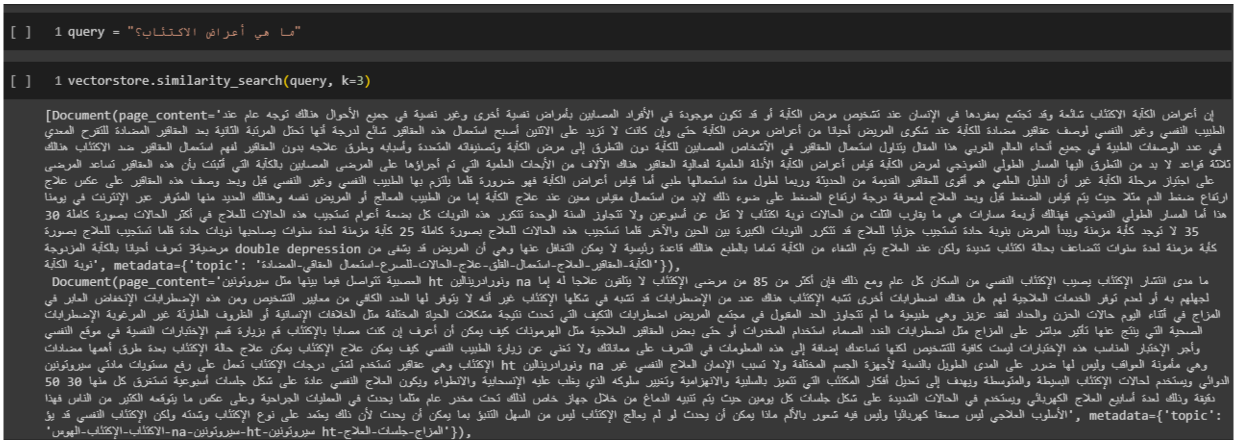
4.3 RAG Chatbot:

The results of the Pinecone similarity search yielded favorable outcomes, indicating the efficacy of the similarity search algorithm in accurately retrieving relevant information from the dataset. The success of the Pinecone similarity search can be attributed to several factors. Firstly, Pinecone utilizes advanced similarity search algorithms optimized for handling high-dimensional embeddings efficiently. By employing state-of-the-art techniques, Pinecone is able to effectively index and search through Nafsy datasets with high-dimensional embeddings, ensuring fast and accurate retrieval of similar items.

Additionally, the quality of the results is indicative of the effectiveness of the embeddings generated by the chosen embedding model (e.g., BERT-based models or Sentence Transformers). These embeddings capture semantic similarities between text items, enabling Pinecone to identify items with similar semantic content to the query.

Furthermore, the success of the Pinecone similarity search underscores the importance of proper parameter tuning and optimization. Parameters such as the dimensionality of the embeddings, the distance metric used for similarity calculation (e.g., cosine similarity), and the threshold for similarity scores can significantly impact the quality of the search results. By carefully selecting and tuning these parameters, Pinecone can deliver highly relevant and accurate results.

Overall, the positive outcomes of the Pinecone similarity search validate its effectiveness in retrieving relevant information from the dataset, affirming its utility as a powerful tool for similarity-based including information retrieval fig ().



```
query = "ما هي أعراض الاكتئاب؟"
```

```
vectorstore.similarity_search(query, k=3)
```

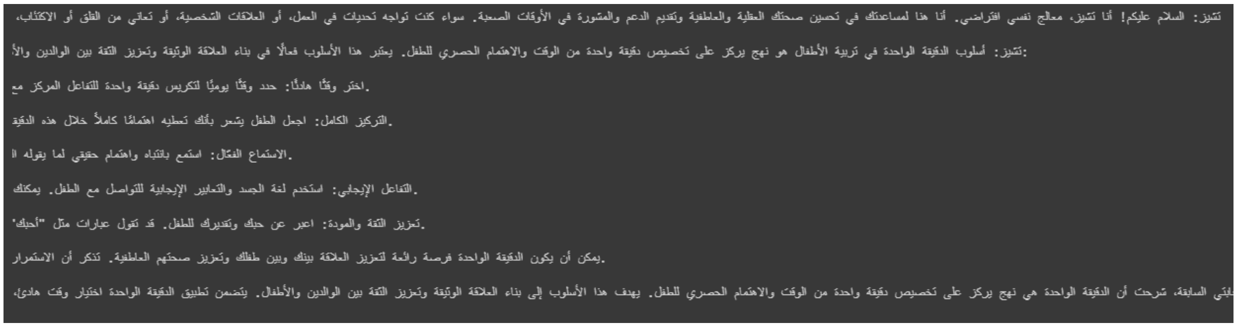
```
[Document(page_content='إن أعراض الكآبة الاكتئاب شائعة وقد تجتمع بمفردها في الإنسان عند تشخيص مرض الكآبة أو قد تكون موجودة في الأفراد المصابين بأمراض نفسية أخرى وغير نفسية في جميع الأحوال هنالك توجه عام عند
الطبيب النفسي وغير النفسي لوصف عقاقير مضادة للكآبة عند شكوى المريض أحيانا من أعراض مرض الكآبة حتى وإن كانت لا تزيد على الاثنين أصبح استعمال هذه العقاقير شائع لدرجة أنها تحتل المرتبة الثانية بعد العقاقير المضادة للتقرح المعدي
في عدد الوصفات الطبية في جميع أنحاء العالم الغربي هذا المقال يتناول استعمال العقاقير في الآشخاص المصابين للكآبة دون التطرق إلى مرض الكآبة وتصنيفاته المتعددة وأسبابه وطرق علاجه بدون العقاقير لفهم استعمال العقاقير ضد الاكتئاب هنالك
ثلاثة قواعد لا بد من التطرق اليها المسار الطولي النموذجي لمرض الكآبة قياس أعراض الكآبة الأدلة العلمية لفعالية العقاقير هناك الآلاف من الأبحاث العلمية التي تم أجراؤها على المرضى المصابين بالكآبة التي أثبتت بأن هذه العقاقير تساعد المرضى
على اجتياز مرحلة الكآبة غير أن الدليل العلمي هو أقوى للعقاقير القديمة من الحديثة وربما لطول مدة استعمالها طبي أما قياس أعراض الكآبة فهو ضرورة قلما يلتزم بها الطبيب النفسي وغير النفسي قبل وبعد وصف هذه العقاقير على عكس علاج
ارتفاع ضغط الدم مثلا حيث يتم قياس الضغط قبل وبعد العلاج لمعرفة درجة ارتفاع الضغط على ضوء ذلك لابد من استعمال مقياس معين عند علاج الكآبة إما من الطبيب المعالج أو المريض نفسه وهنالك العديد منها المتوفر عبر الإنترنت في يومنا
هذا أما المسار الطولي النموذجي فهنالك أربعة مسارات هي ما يقارب الثلث من الحالات نوبة اكتئاب لا تقل عن أسبوعين ولا تتجاوز السنة الوحدة تتكرر هذه النوبات كل بضعة أعوام تستجيب هذه الحالات للعلاج في أكثر الحالات بصورة كاملة 30
35 لا توجد كأبة مزمنة ويبدأ المرض بنوبة حادة تستجيب جزئيا للعلاج قد تتكرر النوبات الكبيرة بين الحين والآخر قلما تستجيب هذه الحالات للعلاج بصورة كاملة 25 كآبة مزمنة لعدة سنوات يصاحبها نوبات حادة قلما تستجيب للعلاج بصورة
كأبة مزمنة لعدة سنوات تتضاعف بحالة اكتئاب شديدة ولكن عند العلاج يتم الشفاء من الكآبة تماما بالطبع هنالك قاعدة رئيسية لا يمكن التغافل عنها وهي أن المريض قد يشفى من double depression مرضية3 تعرف أحيانا بالكآبة المزدوجة
نوبة الكآبة', metadata={'topic': 'الكآبة-العقاقير-العلاج-استعمال-القلق-علاج-الحالات-للصرع-استعمال العقاقي-المضادة'}),
 Document(page_content='ما مدى انتشار الإكتئاب يصيب الإكتئاب النفسي من السكان كل علم ومع ذلك فإن أكثر من 85 من مرضى الإكتئاب لا يتلقون علاجا له إما na ونورادرينالين ht العصبية تتواصل فيما بينها مثل سيروتونين
لجهلهم به أو لعدم توفر الخدمات العلاجية لهم هل هناك اضطرابات أخرى تشبه الإكتئاب هناك عدد من الإضطرابات قد تشبه في شكلها الإكتئاب غير أنه لا يتوفر لها العدد الكافي من معايير التشخيص ومن هذه الإضطرابات الإنخفاض العابر في
المزاج في أثناء اليوم حالات الحزن والحداد لفقد عزيز وهي طبيعية ما لم تتجاوز الحد المقبول في مجتمع المريض اضطرابات التكيف التي تحدث نتيجة مشكلات الحياة المختلفة مثل الخلافات الإنسانية أو الظروف الطارئة غير المرغوبة الإضطرابات
الصحية التي ينتج عنها تأثير مباشر على المزاج مثل اضطرابات الغدد الصماء استخدام المخدرات أو حتى بعض العقاقير العلاجية مثل الهرمونات كيف يمكن أن أعرف إن كنت مصابا بالإكتئاب قم بزيارة قسم الإختبارات النفسية في موقع النفسي
وأجر الإختبار المناسب هذه الإختبارات ليست كافية للتشخيص لكنها تساعدك إضافة إلى هذه المعلومات في التعرف على معاناتك ولا تغني عن زيارة الطبيب النفسي كيف يمكن علاج الإكتئاب يمكن علاج حالة الإكتئاب بعدة طرق أهمها مضادات
وهي مأمونة العواقب وليس لها ضرر على المدى الطويل بالنسبة لأجهزة الجسم المختلفة ولا تسبب الإدمان العلاج النفسي غير na ونورادرينالين ht الإكتئاب وهي عقاقير تستخدم لشتى درجات الإكتئاب تعمل على رفع مستويات مادتي سيروتونين
الدوائي ويستخدم لحالات الإكتئاب البسيطة والمتوسطة ويهدف إلى تعديل أفكار المكتئب التي تتميز بالسلبية والانهزامية وتغيير سلوكه الذي يغلب عليه الإنسحابية والانطواء ويكون العلاج النفسي عادة على شكل جلسات أسبوعية تستغرق كل منها 30 50
دقيقة وذلك لعدة أسابيع العلاج الكهربائي ويستخدم في الحالات الشديدة على شكل جلسات كل يومين حيث يتم تنبيه الدماغ من خلال جهاز خاص لذلك تحت مخدر عام مثلما يحدث في العمليات الجراحية وعلى عكس ما يتوقعه الكثير من الناس فهذا
الأسلوب العلاجي ليس صعقا كهربائيا وليس فيه شعور بالألم ماذا يمكن أن يحدث لو لم يعالج الإكتئاب ليس من السهل التنبؤ بما يمكن أن يحدث لأن ذلك يعتمد على نوع الإكتئاب وشدته ولكن الإكتئاب النفسي قد يؤ', metadata={'topic':
'المزاج-جلسات-العلاج-ht سيروتونين-ht-سيروتونين-na-الاكتئاب-الإكتئاب-الهوس'}),
```

Figure (18): Pinecone Similartiy Search Results.

The results obtained from the Chatbot with the Retrieval-Augmented Generation (RAG) model demonstrated superior performance compared to the results obtained without it. The term "better" signifies an improvement in various aspects such as response quality, relevance, coherence, and user satisfaction. Several factors contribute to the enhanced performance of the Chatbot with the RAG model. Firstly, the RAG model leverages advanced language models such as GPT-3.5-turbo to generate responses based on the provided schema or context. These language models excel in understanding natural language inputs and generating human-like responses, leading to more coherent and contextually relevant interactions with users.

The incorporation of retrieval-based methods in the RAG model enables the chatbot to retrieve and incorporate relevant information from the knowledge base into the generated responses. Additionally, the use of the RAG model facilitates a more dynamic and engaging conversational experience for users. By generating responses that are tailored to the context of the conversation and incorporating relevant information from external sources, the chatbot with the RAG model can maintain more meaningful and interactive dialogues with users, leading to higher levels of user satisfaction and engagement.

Overall, the results obtained from the Chatbot with the RAG model demonstrate its effectiveness in improving the quality and performance of the chatbot system compared to the results obtained without it. This highlights the value of incorporating advanced language models and retrieval-based methods in chatbot systems to enhance their capabilities and deliver more compelling user experiences fig (19).



تشيز: السلام عليكم! أنا تشيز، معالج نفسي افتراضي. أنا هنا لمساعدتك في تحسين صحتك العقلية والعاطفية وتقديم الدعم والمشورة في الأوقات الصعبة. سواء كنت تواجه تحديات في العمل، أو العلاقات الشخصية، أو تعاني من القلق أو الاكتئاب،

:تشيز: أسلوب الدقيقة الواحدة في تربية الأطفال هو نهج يركز على تخصيص دقيقة واحدة من الوقت والاهتمام الحصري للطفل. يعتبر هذا الأسلوب فعالًا في بناء العلاقة الوثيقة وتعزيز الثقة بين الوالدين والأ

.اختر وقتًا هادئًا: حدد وقتًا يوميًا لتكريس دقيقة واحدة للتفاعل المركز مع

.التركيز الكامل: اجعل الطفل يشعر بأنك تعطيه اهتمامًا كاملاً خلال هذه الدقيق

.الاستماع الفعّال: استمع بانتباه واهتمام حقيقي لما يقوله ال

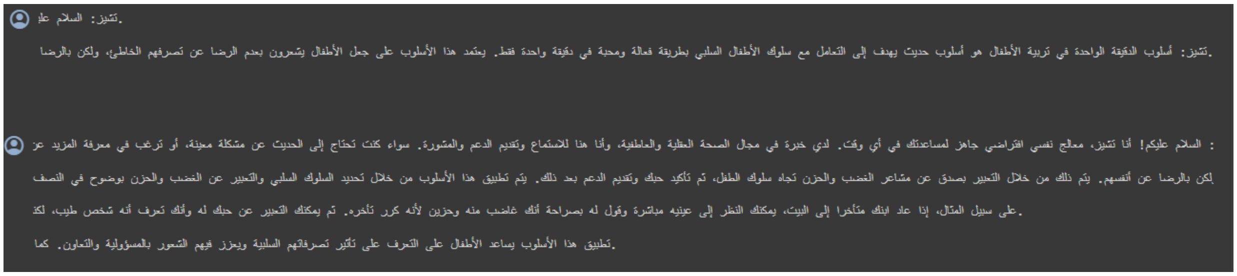
.التفاعل الإيجابي: استخدم لغة الجسد والتعابير الإيجابية للتواصل مع الطفل. يمكنك

.تعزيز الثقة والمودة: اعبر عن حبك وتقديرك للطفل. قد تقول عبارات مثل "أحبك"

.يمكن أن يكون الدقيقة الواحدة فرصة رائعة لتعزيز العلاقة بينك وبين طفلك وتعزيز صحتهم العاطفية. تذكر أن الاستمرار

ابتي السابقة، شرحت أن الدقيقة الواحدة هي نهج يركز على تخصيص دقيقة واحدة من الوقت والاهتمام الحصري للطفل. يهدف هذا الأسلوب إلى بناء العلاقة الوثيقة وتعزيز الثقة بين الوالدين والأطفال. يتضمن تطبيق الدقيقة الواحدة اختيار وقت هادئ،

Figure (19): Chatting with Chezlong without RAG.



.تشيز: السلام عل

.تشيز: أسلوب الدقيقة الواحدة في تربية الأطفال هو أسلوب حديث يهدف إلى التعامل مع سلوك الأطفال السلبي بطريقة فعالة ومحبة في دقيقة واحدة فقط. يعتمد هذا الأسلوب على جعل الأطفال يشعرون بعدم الرضا عن تصرفهم الخاطئ، ولكن بالرضا

: السلام عليكم! أنا تشيز، معالج نفسي افتراضي جاهز لمساعدتك في أي وقت. لدي خبرة في مجال الصحة العقلية والعاطفية، وأنا هنا للاستماع وتقديم الدعم والمشورة. سواء كنت تحتاج إلى الحديث عن مشكلة معينة، أو ترغب في معرفة المزيد عز

لكن بالرضا عن أنفسهم. يتم ذلك من خلال التعبير بصدق عن مشاعر الغضب والحزن تجاه سلوك الطفل، ثم تأكيد حبك وتقديم الدعم بعد ذلك. يتم تطبيق هذا الأسلوب من خلال تحديد السلوك السلبي والتعبير عن الغضب والحزن بوضوح في النصف

.على سبيل المثال، إذا عاد ابنك متأخرا إلى البيت، يمكنك النظر إلى عينيه مباشرة وقول له بصراحة أنك غاضب منه وحزين لأنه كرر تأخره. ثم يمكنك التعبير عن حبك له وأنك تعرف أنه شخص طيب، لكذ

.تطبيق هذا الأسلوب يساعد الأطفال على التعرف على تأثير تصرفاتهم السلبية ويعزز فيهم الشعور بالمسؤولية والتعاون. كما

Figure (20): Chatting with Chezlong with RAG and Nafsy Dataset.

Chapter 5

Conclusion

5.1 Summary of Findings:

In summary, the project has delivered promising outcomes in the development of a sophisticated chatbot system for mental health support in Arabic. The main findings of the project highlight the effectiveness of various technologies and methodologies employed in enhancing the chatbot's capabilities and performance.

Key findings include the successful implementation of advanced natural language processing techniques, such as tokenization, embedding, and topic modeling, which have enabled the chatbot to understand and generate contextually relevant responses. The integration of state-of-the-art language models, retrieval-based methods, and memory capabilities has further enhanced the chatbot's conversational depth, coherence, and engagement.

Additionally, the project has demonstrated the utility of external resources and services, such as Pinecone for similarity search, Faiss for vector storage, and Hugging Face models for text representation, in augmenting the chatbot's functionality and effectiveness. The utilization of these resources has enabled the chatbot to provide accurate information, personalized support, and relevant recommendations to users, contributing to a more enriching and meaningful user experience.

Overall, the project's accomplishments signify a significant advancement in the development of AI-powered Arbic chatbot systems for mental health support. By leveraging cutting-edge technologies and methodologies, the project has laid the foundation for scalable and efficient chatbot solutions that can effectively assist individuals in managing their mental well-being and accessing support resources.

5.2 Limitations and Future Work:

Despite the achievements and advancements made in the development of the chatbot system, it is important to acknowledge certain limitations that should be considered in future iterations and implementations. From one perspective, it is essential to continually update and diversify the training data to ensure that the chatbot remains inclusive, accurate, and sensitive to the diverse needs and experiences of its users.

Furthermore, while the chatbot system incorporates memory capabilities to remember previous conversations, the depth and scope of its memory may be limited, potentially affecting its ability to maintain continuity and coherence in extended dialogues or complex interactions. Enhancements in memory management and retention mechanisms may be necessary to address this limitation and improve the chatbot's ability to recall and utilize relevant information from past interactions effectively.

Additionally, the chatbot system may encounter challenges related to user privacy, data security, and ethical considerations, particularly in handling sensitive and personal information related to mental health. Ensuring compliance with relevant privacy regulations and implementing robust security measures to safeguard user data are critical aspects that require ongoing attention and diligence in the development and deployment of the chatbot system.

Lastly, the chatbot system's performance and effectiveness may be influenced by external factors such as internet connectivity, device compatibility, and user interface design. Addressing these technical constraints and optimizing the chatbot's accessibility, usability, and performance across diverse platforms and environments will be essential to maximize its reach and impact.

One significant limitation encountered in developing the chatbot system for mental health support is the scarcity of Arabic language resources tailored for natural language processing (NLP). This scarcity restricts access to large-scale, high-quality training datasets, pre-trained language models, and standardized evaluation benchmarks specific to Arabic text. As a result, the chatbot's ability to understand, generate accurate responses, and perform effectively in Arabic is hindered. Addressing this limitation requires collaborative efforts to create and curate Arabic language resources, develop specialized NLP models, and establish standardized evaluation metrics, ultimately enhancing the chatbot's effectiveness for Arabic-speaking users.

Another limitation faced is the constraint of computational resources, which can affect the scalability and performance of the chatbot system. Future work may focus on optimizing resource-intensive processes, such as training and inference of language models, by leveraging distributed computing techniques or cloud-based solutions. Additionally, exploring techniques for model compression, quantization, or efficient model architectures could help mitigate computational resource constraints without sacrificing performance.

In conclusion, while the chatbot system for mental health support has demonstrated promising capabilities and potential benefits, it is important to recognize and address the aforementioned limitations to ensure its continued effectiveness, reliability, and ethical integrity in supporting individuals' mental well-being. Ongoing research, development, and collaboration efforts will be crucial in overcoming these challenges and advancing the field of AI-powered mental health support.

References

1. G. Caldarini, S. Jaf, and K. McGarry, “A Literature Survey of Recent Advances in Chatbots,” Information, vol. 13, no. 1, p. 41, Jan. 2022, doi: 10.3390/info13010041
2. Codecademy, “History of Chatbots,” Codecademy, 2013. https://www.codecademy.com/article/history-of-chatbots (accessed Feb. 19, 2023).
3. “A Brief History of Chatbots,” Perception, Control, Cognition, Mar. 26, 2018. https://pcc.cs.byu.edu/2018/03/26/a-brief-history-of-chatbots/ (accessed Feb. 19, 2023).
4. Wikipedia Contributors, “ChatGPT,” Wikipedia, Feb. 19, 2023. https://en.wikipedia.org/wiki/ChatGPT (accessed Feb. 19, 2023).
5. “Technology and the Future of Mental Health Treatment,” National Institute of Mental Health (NIMH), 2015. https://www.nimh.nih.gov/health/topics/technology-and-the-future-of-mental-health-treatment (accessed Feb. 19, 2023).
6. Adnan Asar, “Council Post: Five Tech Innovations That Changed Mental Health In 2020,” Forbes, Apr. 14, 2022. Accessed: Feb. 19, 2023. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2020/11/25/five-tech-innovations-that-changed-mental-health-in-2020/?sh=3215c62c1e9c
7. A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, “Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape,” The Canadian Journal of Psychiatry, vol. 64, no. 7, pp. 456–464, Mar. 2019, doi: https://doi.org/10.1177/0706743719828977.
8. S. D’Alfonso, “AI in mental health,” Current Opinion in Psychology, vol. 36, pp. 112–117, Dec. 2020, doi: https://doi.org/10.1016/j.copsyc.2020.04.005.
9. M. Anastasiia and T. Korotyeva, “A chatbot of a person’s emotional state using a neural network,” 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), Nov. 2022, doi: https://doi.org/10.1109/csit56902.2022.10000558.

10. K. Von Schlegell and O. Abuomar, “Mental Health Frequently Asked Questions Chatbot Powered by Machine Learning,” 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Nov. 2022, doi: https://doi.org/10.1109/iceccme55909.2022.9987745.

11. “Artificial Intelligence Powered Chatbot for Mental Healthcare based on Sentiment Analysis,” Ieee.org, 2023, doi: https://doi.org/10.1109/ICAST55766.2022.10039548.

12. S. Abu Noman Siddik, B. M. Arifuzzaman, and A. Kalam, “Psyche Conversa - A Deep Learning Based Chatbot Framework to Detect Mental Health State,” 2022 10th International Conference on Information and Communication Technology (ICoICT), Aug. 2022, doi: https://doi.org/10.1109/icoict55009.2022.9914844.

13. V. Dhanasekar, Y. Preethi, V. S, P. J. I. R, and B. P. M, “A Chatbot to promote Students Mental Health through Emotion Recognition,” 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Sep. 2021, doi: https://doi.org/10.1109/icirca51532.2021.9544838.

14. F. Patel, R. Thakore, I. Nandwani, and S. K. Bharti, “Combating Depression in Students using an Intelligent ChatBot: A Cognitive Behavioral Therapy,” 2019 IEEE 16th India Council International Conference (INDICON), Dec. 2019, doi: https://doi.org/10.1109/indicon47234.2019.9030346.

15. J. Jia, “The Study of the Application of a Web-Based Chatbot System on the Teaching of Foreign Languages,” 2019. https://www.semanticscholar.org/paper/The-Study-of-the-Application-of-a-Web-Based-Chatbot-Jia/772bfd084867cc19d1e1718dae2805ff3fd4c8fb (accessed Feb. 20, 2023).

16. K. Chung and R. C. Park, “Chatbot-based heathcare service with a knowledge base for cloud computing,” Cluster Computing, pp. 1–13, 2018.

17. Wikipedia Contributors, “Generative pre-trained transformer,” Wikipedia, Feb. 21, 2023. https://en.wikipedia.org/wiki/Generative_pre-trained_transformer (accessed Feb. 21, 2023).
18. Wikipedia Contributors, “Transformer (machine learning model),” Wikipedia, Feb. 19, 2023. https://en.wikipedia.org/wiki/Transformer_(machine_learning_model) (accessed Feb. 21, 2023).
19. A. Vaswani et al., “Attention is All you Need,” Advances in Neural Information Processing Systems, vol. 30, 2017, Accessed: Feb. 21, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
20. A. Radford and K. Narasimhan, “Improving Language Understanding by Generative Pre-Training,” 2018. https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035 (accessed Feb. 21, 2023).
21. Kaggle. "Arabic Psychology Dataset", Kaggle, Available: https://www.kaggle.com/datasets/husamal/arabicphyscologydataset?select=nafsy.csv
22. LangChain Documentation. "Introduction," LangChain. Available: https://python.langchain.com/docs/get_started/introduction
23. LangChain Documentation. "Token Splitter," LangChain. Available: https://js.langchain.com/docs/modules/data_connection/document_transformers/token_splitter
24. Natural Language Toolkit (NLTK) Documentation. (n.d.). Available: https://www.nltk.org/
25. Python Software Foundation. "re - Regular expression operations," Python 3 Documentation. Available: https://docs.python.org/3/library/re.html
26. OpenAI. "BERTopic: Unsupervised Topic Modeling with BERT in Python," GitHub, Available: https://github.com/MaartenGr/BERTopic.
27. Asafaya, "asafaya/bert-base-arabic," Hugging Face.

28. Sentence Transformers. "paraphrase-multilingual-MiniLM-L12-v2," Hugging Face, Available: [https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2].
29. Facebook Research. "Faiss: A library for efficient similarity search," GitHub, Available: [https://github.com/facebookresearch/faiss].
30. Pinecone. "Pinecone - A vector database for building similarity search applications," Available: [https://www.pinecone.io/].
31. OpenAI. "GPT-3.5 Models," OpenAI. Available: https://openai.com/gpt-3.5.