Original Article

Design and Development of Amharic Chatbot for Mental Health Using Bidirectional Long Short-Term Memory

Gashaw D. Wubneh [1,], Getnet T. Askale [2], Michael M. Woldeyohannis [3], Worku A. Degife [1]

1. *Department of Information Systems, College of Informatics, University of Gondar, Gondar, ETH*
2. *Department of Computer Science, College of Informatics, University of Gondar, Gondar, ETH*
3. *School of Information Science, College of Natural and Computation Science, Addis Ababa University, Addia Abeba, ETH*

Received: July 21, 2025 | Review began: August 06, 2025 | Review ended: September 18, 2025 | Published: February 25, 2026

Abstract

Nowadays, people are significantly affected by mental illnesses such as depression and anxiety disorders. These conditions can often be treated successfully if appropriate therapy is provided promptly. However, due to a shortage of mental health professionals, limited resources, economic constraints, and fear of social stigma, many people do not receive sufficient therapy. As a result, providing effective mental health interventions remains challenging. This study addresses this gap by designing and developing an Amharic language mental health chatbot to provide supportive counseling for depression and anxiety. The primary research question is as follows: How can a deep learning-based, retrieval-focused Amharic chatbot be effectively implemented to support mental health care? To achieve this, the study employed a design science research methodology, integrating natural language preprocessing with a Bidirectional Long Short-Term Memory (BiLSTM) network and Word2Vec embeddings to capture semantic relationships. Data were collected from both documented and non-documented sources, preprocessed, and structured into intents, patterns, and responses. Experimental results demonstrate that the BiLSTM model achieved 91.25% accuracy in classifying user inputs. In addition, a preliminary User Acceptance Test involving mental health experts and volunteers yielded an average satisfaction score of 86.6%, confirming that the system is user-friendly, provides clear responses, and is practically applicable in real-world settings. Unlike prior rule-based or English-language chatbots, this work makes a novel contribution by applying advanced deep learning techniques to a low-resource, morphologically complex language in a culturally sensitive domain. The findings highlight the potential of natural language processing and deep learning to deliver scalable, accessible, and stigma-free mental health support in Ethiopia and similar low-resource settings.

Categories: AI/ML-based decision support systems, Deep Learning, Natural Language Processing (NLP)

Keywords: chatbot, deep learning, lstm, bilstm, word2vec, natural language processing, mental health, psychotherapy, depression, anxiety

Introduction

Mental health disorders affect millions of people worldwide and remain a critical public health concern. Alarmingly, one person dies every 40 seconds due to suicide [1]. As technology continues to make communication platforms more accessible, there is a growing need to develop tools that can understand and interact in the languages people use daily [2]. Depression, a disabling mental disorder, is characterized by feelings of worthlessness, sadness, and loss of interest in

Natural Language Processing (NLP) is a branch of artificial intelligence that enables computers to understand, interpret, and generate human language in the form of text or speech [3]. NLP applications include machine translation, speech recognition, speech synthesis, chatbots, and pattern recognition. A chatbot is an intelligent conversational agent that simulates human conversation through text, voice commands, or both. It can initiate and maintain a conversation by interpreting user inputs and generating meaningful responses. The effectiveness of a chatbot depends heavily on the relevance and accuracy of its responses [4]. Depending on the application, the interaction can be textual or auditory. NLP techniques enable chatbots to understand user queries and respond appropriately [5].

Traditionally, patients with mental health concerns rely on direct communication with psychiatrists or psychologists to seek help and treatment [6]. However, patients often face long waiting times due to overloaded health professionals, limited human resources, and insufficient infrastructure to provide continuous support [4]. The shortage of psychotherapy experts, high consultation fees with psychiatrists and counseling psychologists, and limited access to mental health services further exacerbate the challenge [7,8].

To address these issues, this study proposes an innovative Amharic-language chatbot model designed to provide personalized psychotherapy for patients suffering from mental health conditions such as anxiety and depression. The proposed chatbot allows users to express their feelings in text, and the chatbot provides supportive therapeutic responses based on the input. The only requirement for users is to input their feelings or questions related to their mental health in Amharic through a dialogue box. To develop this chatbot, core NLP techniques such as text normalization, tokenization, and stop-word removal are employed during preprocessing, alongside word embedding methods and deep learning techniques like recurrent neural networks with Bidirectional Long Short-Term Memory (BiLSTM) architectures.

Related works

Nowadays, different research is conducted on the area of chatbots for dissimilar purposes. This section provides a comprehensive overview of fundamental works on the mental health chatbot offering psychotherapy. Chatbots were designed either with a rule-based self-learning or a corpus-based approach. In a rule-based chatbot, responses are generated based on a fixed set of handcrafted "if-then" rules and keyword matching. For example, if a user's input contains the word "sad," the rule might trigger a predefined response like "I'm sorry to hear you're feeling sad." These systems are rigid and cannot handle queries outside their predefined rules. In contrast, self-learning chatbots model use machine learning algorithms to learn appropriate responses from data, allowing for more flexible and context-aware interactions. These bots can be of two types: using retrieval-based models and generative-based. For each question in the Retrieval Based Bot, use a repository of predefined answers and some kind of heuristic to select an appropriate answer based on the input. The generative model generates better answers than the other three models, based on current and past user messages [8].

The first chatbot was ELIZA [9], developed by Joseph Weizenbaum to simulate a psychotherapist in clinical treatment. ELIZA simulated conversation by using pattern matching; on the other hand, it used the input and output rules and keywords to create the response for the user. If the input keyword matched with its stored keyword-rules repository, the sentence is mapped (generated response) according to a rule associated with the keyword [10,11]. The main limitation is that if the user input does not match certain conditions, the system cannot generate a response, and ELIZA cannot understand the content of its conversations [11].

ALICE was developed by Dr. Wallace based on natural language understanding and pattern matching. The breakthrough came when NLP was used to solve the question-and-answer task. Before the matching process begins, ALICE applies a normalization process to each input: it removes all punctuation, splits the input into two or more sentences if appropriate, and converts the input to uppercase. The Artificial Intelligence Markup Language interpreter then tries to match word by word to get the longest pattern match, using a simple depth-first search. However, it is implemented using a rule-based approach, which means it does not truly understand user input [12]; the user's input must match the dataset pattern exactly.

Srivastava and Singh [11] developed an automated medical chatbot designed to interact with users through natural language and assist in medical diagnosis. The chatbot was implemented using Artificial Intelligence Markup Language, where user inputs are processed through predefined patterns and templates to generate appropriate responses. The system predicts diseases based on the symptoms provided by users, employing machine learning algorithms such as K-nearest neighbors, support vector machine, and naive bayes, achieving respective accuracies of 94.6%, 88.6%, and 80%. Despite these promising results, the system exhibits limitations in comprehending user input beyond the predefined patterns, indicating that its understanding is constrained to inputs that match the programmed templates.

Another attempt is the implementation of a machine learning-based Bangla healthcare chatbot [12], in a study conducted for Bangla-speaking users. The main aim of this research is to collect the user's symptoms through Bangladeshi text and predict the disease based on those symptoms. Cosine similarity and Term Frequency-Inverse Document Frequency techniques have been applied to measure the similarity between symptoms stored in the knowledge base and the user's input. Based on this, the system predicts the disease using a dataset. To achieve this, the researchers used support vector machine models. However, the system can predict the disease accurately only if the user provides at least three symptoms; otherwise, the prediction is inaccurate, and the features are extracted using handcrafted rules.

A chatbot for psychiatric counseling in mental healthcare services was developed based on emotional dialogue analysis and sentence generation [7]. It provides conversational service for mental health care based on emotion recognition methods and the chat assistant platform. In this study, researchers used natural language understanding to understand counseling contents and emotion recognition based on a multi-modal approach from conversation content, intonation, and facial expression integrated with traditional recurrent neural networks. However, this model lacks the problem of long-term dependencies. All the above studies were done in English and other languages, but not in Amharic. In addition, most of those studies are implemented with rule-based approaches, and the rest have feature extraction limitations. Therefore, this study sets out to answer the following research question: How can a deep learning-based, retrieval-focused Amharic-language chatbot be effectively implemented to provide supportive counseling for depression and anxiety? The key contributions of this work are twofold: (1) the development of the first Amharic-language mental health chatbot using BiLSTM and Word2Vec embeddings, and (2) the design and evaluation of a prototype to demonstrate usability and practical applicability.

Materials And Methods

The Design Science Research (DSR) approach is a problem-solving paradigm that focuses on the creation and evaluation of innovative artifacts (like models, methods, such as our chatbot) to address identified problems [13]. DSR is increasingly recognized in the field of Information Systems for its practical, solution-oriented nature and its structured guidance for evaluation and iterative development [13,14]. We selected DSR as it provides a rigorous framework for 1) identifying a significant real-world problem (the lack of accessible mental health support in Amharic), 2) designing and developing a novel artifact (the BiLSTM-based chatbot) as a solution, 3) demonstrating its utility through implementation and experimentation, and 4) evaluating its performance and effectiveness through metrics and user testing. This methodology ensures our research is not only academically sound but also delivers a practical, validated tool.

Data collection

Data collection is a foundational step in building the chatbot. For this study, both primary and secondary data sources were utilized.

Primary Data

Data was gathered directly from patients through interviews and recorded audio sessions during psychotherapy consultations. These audio recordings were captured using a mobile device and manually transcribed into Amharic text, as the chatbot processes only text inputs.

We collected primary data in the form of interviews and audio. Interview data included the meaning of mental disorders such as depression and anxiety, the symptoms of those disorders, and other related information. One of the challenges we faced during data collection was patient permission. Some clients did not allow a third person to be recorded in the therapy session, while others were voluntary after the expert explained why the researcher needed to record their speech.

How to cite this article:
Wubneh G D, Askale G T, Woldeyohannis M M, et al. (February 25, 2026) Design and Development of Amharic Chatbot for Mental Health Using Bidirectional Long Short-Term Memory. Cureus J Comput Sci 3 : es44389-026-00034-2. DOI https://doi.org/10.7759/s44389-026-00034-2

Audio data: The researcher recorded the voice of the expert and patient when the expert provided psychotherapy to their patients, mostly in the form of dialogue or question-and-answer. We recorded the conversations between the client and counselor and then converted this data into text. These data were collected from Tsibebe Ghion Specialized Hospital, Bahir Dar University, and the University of Gondar Specialized Hospital. We did not have well-organized mental disorder (depression and anxiety) psychotherapy data, so after collecting data from different sources in the form of questions and answers, these data were structured into tags, patterns, and responses. Our dataset was prepared manually using a Notepad++ text editor in JSON format. The JSON format was created manually, i.e., handcrafted, because we did not have well-organized data. The format includes intent, tag, pattern, and response.

Secondary Data

Additional question-and-answer datasets were extracted from books, published papers, and relevant social media platforms [15,16]. Since many sources were in English, translation into Amharic was required. Those case studies were written in English, so it was mandatory to translate them into Amharic. Initially, we gathered data from different sources, which contained a set of messages and their corresponding responses.

The pattern is known as an utterance, which refers to anything the user says or questions asked by the user. For example: “ድብርት ማለት ምን ማለት ነው?” and “የድብርት ምልክቶች ምን ምን ናቸው?” (English: “what is depression?”, “What are the symptoms of depression?”). The entire sentence is the utterance. Intent refers to the user’s interaction with a chatbot or the intention behind each message that the chatbot receives from a particular user (utterance). Therefore, each information tag has its intent with a unique ID. For example, a user says, “የድብርት ምልክቶች ምን ምን ናቸው?” (What are the symptoms of depression?”). The user intends to retrieve the tag number corresponding to the sentence “የድብርት ምልክቶች.” The response includes an answer to the user's question. For example, if the user asks “ድብርት ማለት ምን ማለት ነው?” (English: “What does depression mean?”), the bot retrieves a response related to the user’s question, such as “የድብርት (ድባቴ) የአንድን ሰዉ የመስራት ሀይል በመግታት ፣ ተስፋ በማስቆረጥ እንዲሁም እንቅስቃሴን በመግታት የተሻለ ስሜት ለማግኘት የሚያደርገውን ሂደት ያወሳስባል” (English: “Depression complicates the process of feeling better by inhibiting a person's ability to function, creating feelings of hopelessness, and inhibiting activity”). For the sake of representative tagging, the researcher applied dialogue or direct user communication with the psychologist or psychotherapist.

The utterances were prepared based on differential expressions of the same concept to generate multiple expressions for a single topic. We accepted expressions from different people because we did not have further expressions of topics. By generalizing them, the researcher generated patterns for the dataset.

The dataset has a minimum of three and a maximum of five utterances for the same content. For additional questions, the retrieval of content from those patterns is handled by the created model.

How to cite this article:
Wubneh G D, Askale G T, Woldeyohannis M M, et al. (February 25, 2026) Design and Development of Amharic Chatbot for Mental Health Using Bidirectional Long Short-Term Memory. Cureus J Comput Sci 3 : es44389-026-00034-2. DOI https://doi.org/10.7759/s44389-026-00034-2

ታካሚ: [ወደዚህ ትምህርት ቤት ተዛውሬ ከገበው አንስቶ የግቢው ተማሪ አያወሩኝም ለኔ ቦታ አይሰጡም]

ቴራፒስት: [ልጆቹን ለማነጋገር እና አብሮ ለመሆን ያደረግሽው ነገር]

ታካሚ: [በፊት መጀመርያ ላይ ምሳም አብሮ ለመብላት ሞክሬ ነበር እነሱ ግን ደስተኛ አይደሉም]

ቴራፒስት: [አሁን የከበደሽ ነገር ይሔ ነው]

ታካሚ: [በጣም ይደብራል:: የበፊቱ ትምህርት ቤት በጣም ብዙ ምርጥ ጓደኞች]

ቴራፒስት: [ስለዚህ ይሔ ላንች በጣም ተለውጦብሻል ያለው ሁኔታ]

ታካሚ: [በጣም ነው እንጂ የተለወጠብኝ , አወ እንዴ በጣም ነው እንጂ የተቀየረብኝ]

ቴራፒስት: [አሁን ከልጆቹ ጋር ተግባብቴ ለመኖር ምን እያደረግሽ]

ታካሚ: [አሁን እማ መሞከር ሰልችቶኛልተቀብየዋለው ምሳም ብቻየን ነው የምበለው ከሌለውም ጋር ለመጫወት አልሞከርም]

ቴራፒስት: [ተስፋ እንደቆረጥሽ ንግግርሽ ያሳያል, ተስፋ እንደሌለሽ አነጋገርሽ ያስታውቃል]

ታካሚ: [በትክክል ተስፋ ቆርጫለው,እውነት ነው በጣም ነው ተስፋየ የተሟጠጠው]

ቴራፒስት: [ከት/ቤት ውጭ እራስሽን ለማዝናናት ምን ታደርጊያለሽ]

ታካሚ: [ከቀድሞ ጓደኞቸ ጋር አወራለሁ:: ቲቪ አያለሁ ፣ ብዙም የማደርገው ነገር]

ቴራፒስት: [በፊትስ ?, ቀድሞስ?]

ታካሚ: [ኳስ እጫወታለው ፣ ድብብቆሽ እጫወታለው፣ ከጓደኞቸ ጋርም ፊልም አያለው]

ቴራፒስት: [በዚህ ሳምንት ያስደበረሽን አጋጣሚ በደንብ ልትነግሪኝ ትችያለሽ]

ታካሚ: [ትናት ቴቪ እያየው ጓደኞቸ ናፈቁኝ እና እንደቀድሞው ጥሩ ነገር እንደማይሰማኝ አሰብኩ]

ቴራፒስት: [እኔም እንደሱ ባስብ በጣም ይደብረኛ:: ከሀሳቡ በኋላ ግን ምን ማድረግ ሞከርሽ?]

FIGURE 1: Secondary source sample data

Figure *1* shows a sample of the secondary data used in our study, presented after being translated from English in a psychopathology case.

Dataset description

The dataset used in this study was curated from both primary and secondary sources. The primary data consisted of audio recordings of psychotherapy sessions in Amharic, transcribed manually into English. Secondary data was collected from published books, academic papers, and online forums on mental health, initially in the English language. English-language sources were translated into Amharic by a certified language specialist, who worked independently to reduce bias. His translations were subsequently cross-checked and validated by a third bilingual expert, a clinical psychotherapist, to ensure both linguistic accuracy and clinical relevance. The resulting dataset is valuable for this initial study, but its hand-curated nature and limited size are recognized as limitations that may affect the generalizability of the interview patterns in real-world contexts.

The final dataset comprises 4,450 utterances labeled into 1,004 intent categories (tags), each associated with multiple patterns and responses, as shown in Table *1*. The data were structured in JSON format, with each entry containing fields for tag, patterns, and responses.

Metrics	Value
Total number of intents/tags	1,004
Total utterances	4,450
Average utterances per intent	4
Training data shape (X_train)	3,560
Testing data shape (X_test)	890

TABLE 1: Dataset statistics

Table 1 presents key statistics for the dataset, summarizing the total number of utterances, intents, and the distribution of data for training and testing.

Model design

The proposed system architecture for the Amharic mental health chatbot includes multiple processing phases, starting from user input to generating an appropriate response. The entire methodology we followed is depicted in Figure 2.

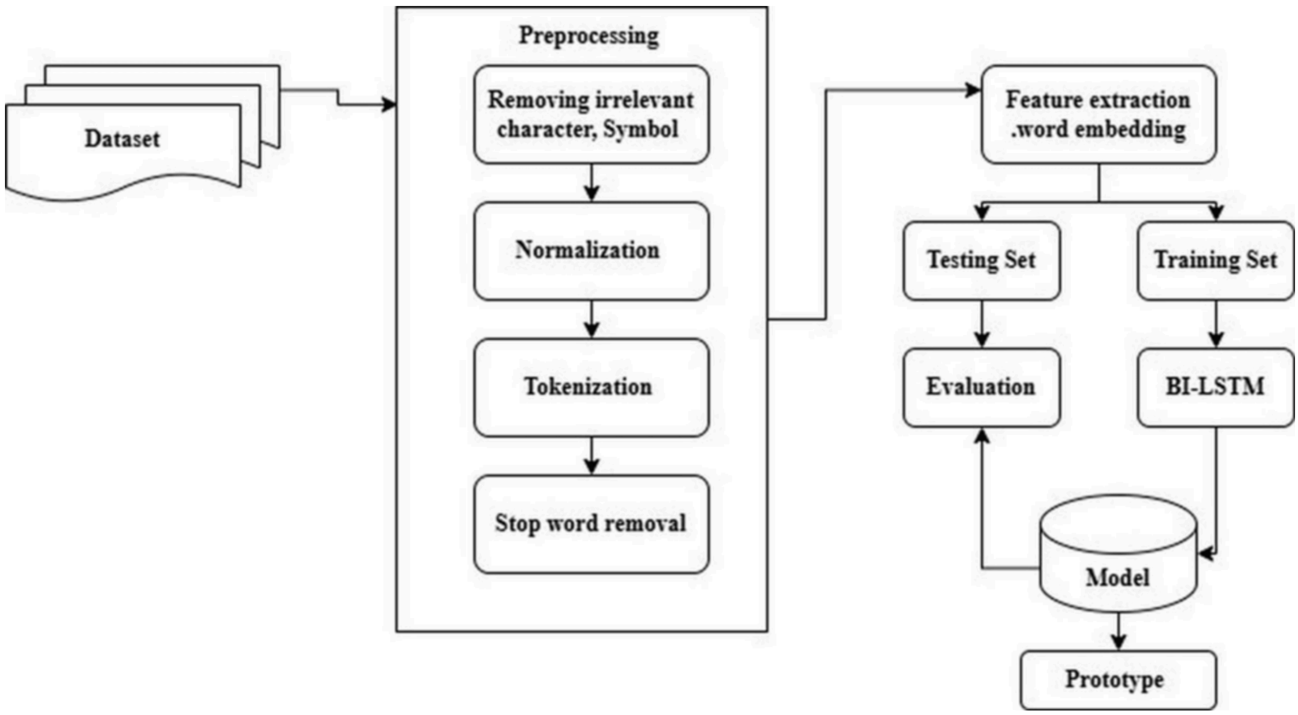

FIGURE 2: Proposed system architecture

Text Cleaning

All irrelevant symbols, non-Amharic characters (such as English letters), punctuation marks, and whitespace characters were removed to standardize the dataset. This step is crucial for minimizing noise and inconsistencies.

Related Works

Normalization is a subprocess that converts letters in documents into a same character. This task is also performed during query analysis and enables the system to be case-insensitive. In Amharic, case sensitivity exists, meaning people may use different characters for the same purpose. The same word written with different letters but the same sound is considered a different word by the retrieval system. Normalization ensures uniform representation of Amharic letters that have similar pronunciations but are written differently, such as ሀ (he), ሐ (hhe), and ኀ (hhhe). For example, both ኀ and ሐ were normalized to ሀ (he). This enhances consistency during training.

Tokenization

Tokenization is the process of breaking raw text into smaller units, such as individual words or sequences of words, known as tokens [17]. These tokens are essential for understanding the context of the text and for developing robust NLP models. The main goal of tokenization is to interpret the meaning of text by analyzing word sequences, thereby making the content suitable for further processing and analysis [17]. In this study, the *word_tokenize* method from the NLTK library was used to segment sentences into individual words or tokens.

Stop Word Removal

Stop word removal is the process of eliminating words that do not add significant meaning to a text corpus. It is well established that stop words, such as common prepositions, articles, and pronouns, do not contribute meaningfully to the context or content of text documents. Removing these words helps reduce the dimensionality of the data, making the text more focused and computationally efficient to process [18]. In this study, an Amharic stop word list provided by another researcher was used to filter out such words from the dataset.

Feature Extraction (Word Embedding)

After all preprocessing steps, the next phase is text vectorization, also known as word embedding, where each text pattern is converted into a numerical vector representation. This process enables the model to capture the semantic similarity between words. For this study, the widely used Word2Vec method was implemented, which is known for generating distributed vector representations of words and effectively capturing contextual word-to-word relationships [19,20].

Before feeding the data into the deep learning model, the words must be converted into feature vectors, since raw text cannot be directly processed by deep learning algorithms. The embedding layer plays a crucial role in this conversion: it transforms each word into a dense vector representation. Each word is assigned a unique integer index, which the embedding layer maps into a fixed-dimensional vector space. To build the Word2Vec model, a custom dataset was prepared and thoroughly preprocessed.

To develop our Word2Vec model, we first trained it with Gensim using our dataset. The Gensim library was used to train the Word2Vec model. First, the vocabulary was built from the tokenized sentences, and then the model was trained to generate word embeddings. The resulting vectors represent words in a multidimensional space where semantically similar words are positioned closer together.

The following parameters were used:

Sentences: The data or tokens that the model is trained on to generate word embeddings. The total number of tokens in our corpus is 9,090, represented as a list of lists.

Embedding size = 100: Each word is represented by an index, and each index is converted into a 100-dimensional vector.

Window = 15: The maximum distance between the current word and its adjacent words. Words occurring more than 15 positions away are not considered related to the current word.

min_count = 1: The model considers all words that appear in the dataset, even those appearing only once. This was necessary because our dataset is small.

Workers = 3: This parameter accelerates training by executing multiple threads simultaneously.

Negative sampling = 10: Used to further optimize training efficiency.

After setting these parameters, the vocabulary was built from a sequence of sentences, and the model was initialized. This step involved processing all words, filtering unique words, and performing basic frequency counts.

The trained model effectively identifies relationships between words within a sentence by calculating word similarity based on context and distance, and by representing similar words with comparable vector dimensions. Words that appear in similar contexts are embedded in close proximity within the vector space, indicating their semantic closeness. The model calculates word similarity by analyzing the distance between vectors, thereby grouping semantically similar words together. Our model finds similarity at the sentence level.

Classification model

The dataset was split into 80% training data and 20% testing data. The training data were used to train the model, while the testing data were used to evaluate its accuracy. The training data were then fed into a BiLSTM network to build the model. To classify user questions, the neural network takes preprocessed input and processes it through four layers: an embedding layer, which converts words into dense vector representations; a BiLSTM layer, which is capable of learning sequential dependencies in both forward and backward directions; a dropout layer, which serves as a regularization technique to prevent overfitting; and a dense output layer with a softmax activation function for classification.

After training the proposed model on the training dataset, the model was saved. For testing, the saved model was loaded and evaluated using new inputs. Finally, model performance was assessed using the testing dataset along with the developed prototype.

Results And Discussion

This section presents the detailed implementation and experimental evaluation of the proposed Amharic mental health chatbot system. It covers the development environment, dataset sources, and the preprocessing techniques applied to prepare the data. Furthermore, it explains the word embedding process, dataset-splitting strategy, and the architecture used to build the deep learning model, particularly the BiLSTM network. It also outlines the integration of the trained model into a functional prototype and provides a comprehensive analysis of the experimental results, including model performance metrics and user acceptance testing (UAT). All aspects are discussed with the necessary scientific rigor to validate the effectiveness and practicality of the proposed system.

Experimental setup

To evaluate the performance of the proposed Amharic mental health chatbot, we conducted experiments using the following environment:

· Hardware: Intel Core i7 @ 2.80GHz, 16 GB RAM, NVIDIA GeForce MX450 GPU

· Operating System: Windows 10 (64-bit)

· Programming Language: Python 3.9

· Development Tools: Jupyter Notebook, Tkinter for GUI prototype

· Libraries Used: TensorFlow, Keras, NLTK, NumPy, Pandas, Scikit-learn, Word2Vec

Data was collected from both documented sources (books, case records) and non-documented sources (interviews with psychiatrists and therapists from the University of Gondar and Bahir Dar University). Preprocessing involved character cleaning, normalization, stop word removal, tokenization, and vector representation using Word2Vec.

The dataset consisted of 3,560 utterances related to depression and anxiety for training, collected from domain experts, psychotherapy documents, and social media content. After preprocessing - including normalization, tokenization, stop word removal, and embedding with Word2Vec - the dataset was split into 80% for training and 20% for testing. The system was trained for 80 epochs using categorical cross-entropy loss and the Adam optimizer, and it was evaluated using accuracy and loss metrics.

The core architecture of the proposed chatbot model is based on a BiLSTM neural network, which is well-suited for capturing contextual information from sequential text data. The model begins with an embedding layer that transforms input tokens into dense vector representations, preserving semantic relationships between words. This embedding layer was initialized using Word2Vec and configured with an input vocabulary size of 9,090, an output embedding dimension of 100, and a maximum sequence length of 100.

Next, a BiLSTM layer with 128 units in each direction processes the input sequences both forward and backward, thereby enhancing context understanding. To mitigate overfitting during training, a dropout layer with a rate of 0.3 is applied. A dense hidden layer with 64 units and ReLU activation further refines the learned features. Finally, the output layer is a Softmax layer with 1,004 units, corresponding to the predefined intent categories. This architecture enables the model to classify user utterances accurately and generate appropriate psychotherapy-related responses based on the learned patterns.

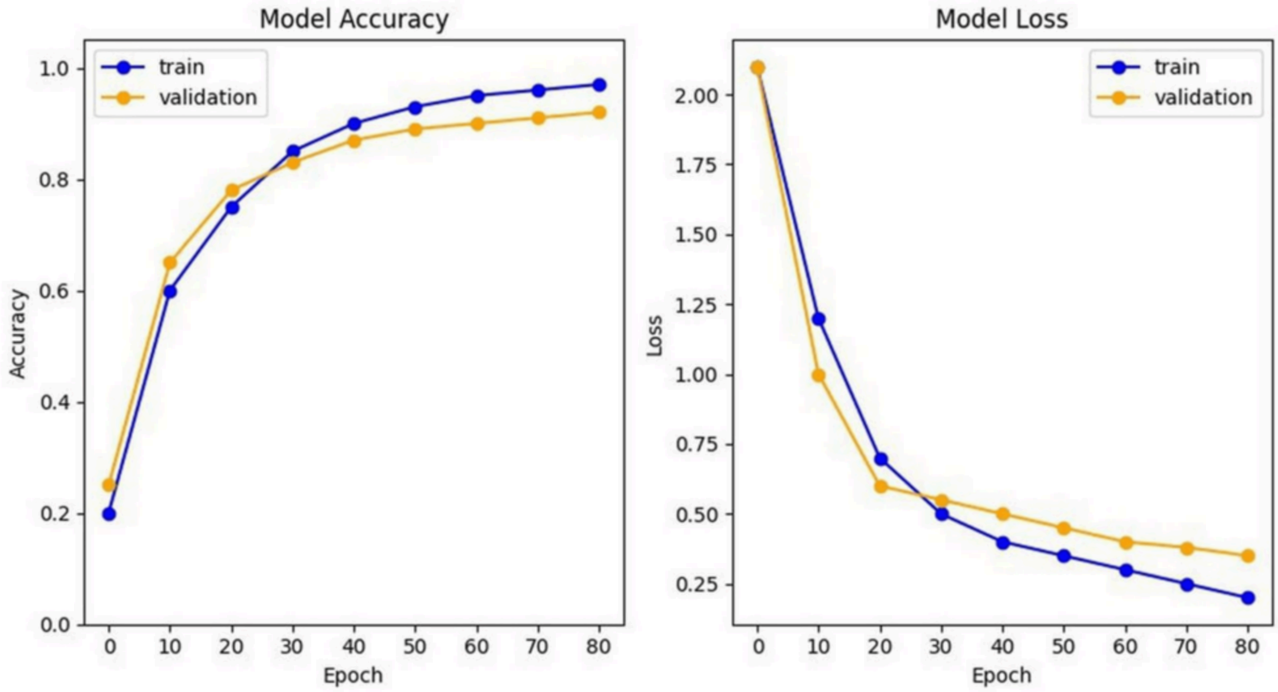

FIGURE 3: Accuracy and loss of the proposed model

Figure *3* presents the accuracy and loss curves of the BiLSTM model during training and testing over 80 epochs. The model achieved a testing accuracy of 91.25%, indicating strong generalization capability.

The learning curves show smooth convergence, with both training and testing losses decreasing steadily across epochs. The absence of significant overfitting suggests that the applied regularization techniques were effective in promoting generalization.

The accuracy curves for training and testing demonstrate a steady increase during the early epochs before gradually plateauing, indicating that the model was learning effectively and reached convergence. The close alignment between training and testing accuracy throughout the process further confirms that the model generalizes well without overfitting.

How to cite this article:
Wubneh G D, Askale G T, Woldeyohannis M M, et al. (February 25, 2026) Design and Development of Amharic Chatbot for Mental Health Using Bidirectional Long Short-Term Memory. Cureus J Comput Sci 3 : es44389-026-00034-2. DOI https://doi.org/10.7759/s44389-026-00034-2

Similarly, the loss curves reveal a consistent decline in both training and testing loss values, with only a small gap between them. This stability reflects the robustness of the model. The final testing accuracy of approximately 91.25% underscores the model's strong performance in classifying user utterances into the correct intent categories.

Prototype evaluation

To evaluate user satisfaction, a User Acceptance Test was conducted with 15 participants, including mental health experts and volunteers. The prototype was assessed based on five criteria using a 5-point Likert scale (1 = Poor, 2 = Fair, 3 = Good, 4 = Very Good, 5 = Excellent). Although the number of participants is relatively small and therefore insufficient for generalization, this limitation was due to the scarcity of available professionals. Despite this, the results, presented in Table 2, indicate a high level of user satisfaction.

No.	Evaluation criteria	Poor	Fair	Good	Very good	Excellent	Average
1	Is the prototype user-friendly to end-users? How easy is it to use and interact with?			3	6	6	4.2
2	Does the prototype respond with clear language?			1	5	9	4.53
3	Does the prototype have a better response?		1	2	6	6	4.13
4	Does the prototype have a better response time?			3	5	7	4.26
5	How do you explain the importance of this prototype?				7	8	4.53
Average performance							4.33

TABLE 2: User acceptance test results

Table 1 shows that the average UAT score is 4.33 out of 5. This corresponds to an overall acceptance rating of 86.6%, indicating that the system is highly usable and effective in providing psychotherapy support. Based on these results, the researchers conclude that the average performance across each criterion contributes positively to the overall evaluation of the prototype.

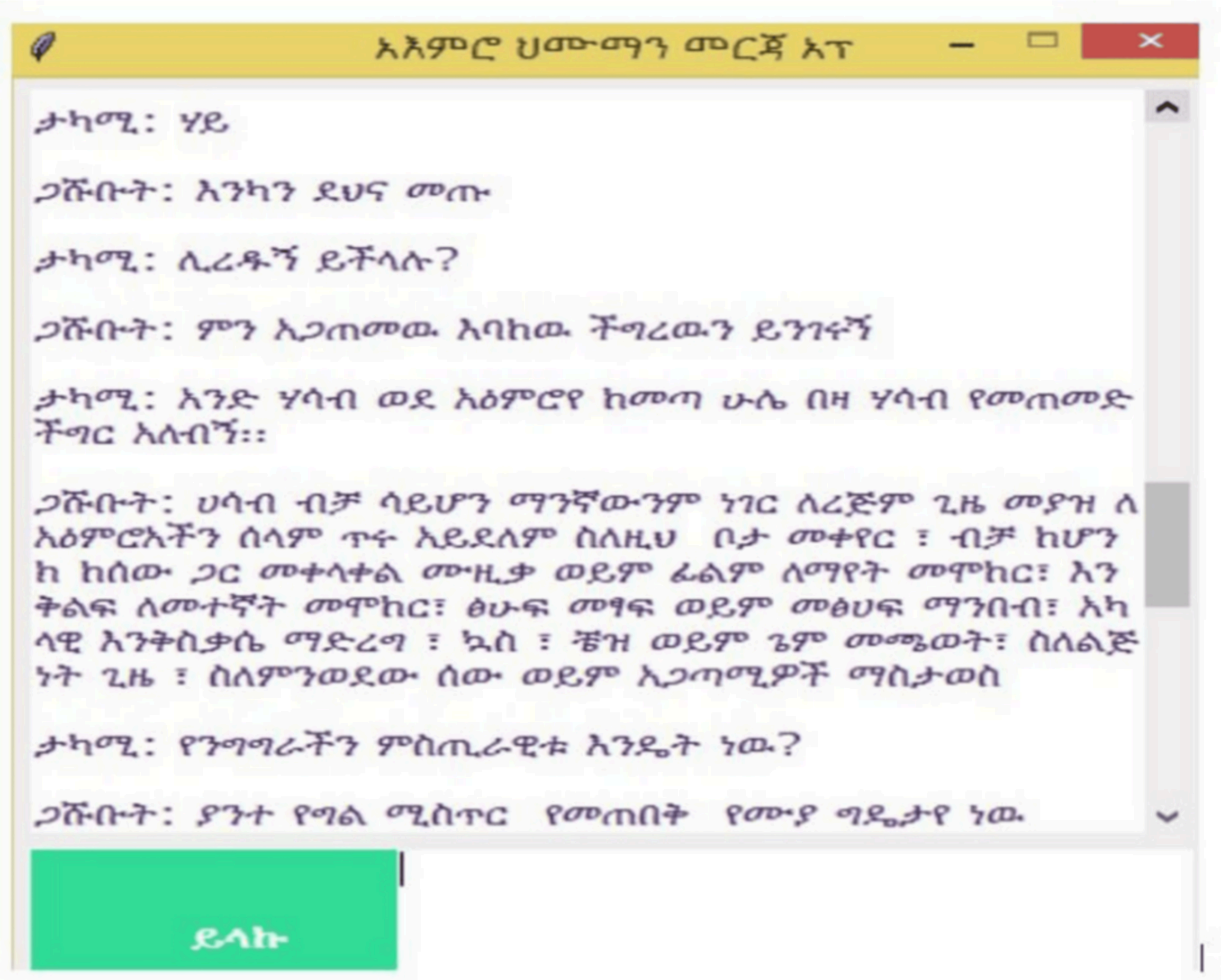

FIGURE 4: Chat interaction between the user and the chatbot prototype

Figure *4* illustrates a sample chat interaction between a user and the developed Amharic mental health chatbot prototype. The interface enables users to input their concerns in Amharic - particularly related to depression and anxiety - and receive automated, contextually appropriate responses.

The chatbot model was evaluated through a series of experiments designed to assess its effectiveness in classifying user inputs and generating relevant counseling responses. The BiLSTM model achieved a final testing accuracy of 91.25%, confirming its strong capability in understanding user intent in a low-resource language setting.

Compared with traditional rule-based systems such as ELIZA [9] and ALICE [12,13], the proposed model offers distinct advantages. ELIZA relied on shallow parsing and keyword substitution, while ALICE employed Artificial Intelligence Markup Language to match inputs with predefined responses. Both approaches lacked flexibility and adaptability, making them unsuitable for emotionally sensitive contexts such as psychotherapy. By contrast, the BiLSTM architecture captures long-range dependencies in sequential data [21-23], which is particularly beneficial for Amharic due to its rich morphology and complex syntactic structures.

The integration of Word2Vec embeddings further enhanced performance by enabling the model to recognize semantic relationships between words [19-22]. Unlike one-hot encoding or Term Frequency-Inverse Document Frequency, Word2Vec preserves contextual meaning, allowing the system to generate appropriate responses even when users phrase their concerns differently. This semantic sensitivity is critical in mental health, where emotional expression can vary widely.

The system also outperformed several existing deep learning chatbot studies focused on English datasets. For instance, models developed by Patel et al. [24] and Sharma et al. [2] combined NLP with traditional classifiers but lacked deep contextual understanding and were not adapted to non-English languages. Moreover, many prior studies overlooked psychotherapy-specific intent categories. In contrast, the dataset in this study was constructed and labeled with the guidance of Ethiopian mental health professionals, ensuring greater domain relevance.

UAT further validated the system's practical utility. With an overall acceptance rate of 86.6%, participants - including psychologists and end users - rated the chatbot highly in terms of user-friendliness, response time, clarity, and usability. These findings suggest strong real-world potential to reduce the burden on mental health professionals by providing immediate, private, and accessible psychological support. The text-based format also helps mitigate cultural stigma around seeking therapy in Ethiopia, as users can engage with the system anonymously.

From an architectural perspective, the BiLSTM offered clear improvements over unidirectional LSTMs. While standard LSTMs process input sequences in a single direction, BiLSTMs analyze both forward and backward contexts, yielding deeper comprehension of user queries. This makes the model especially suitable for Amharic, where critical syntactic and semantic cues may appear at different points within a sentence.

Despite these promising results, several limitations remain. The dataset, though domain-specific, is relatively small and manually labeled. Expanding it with more diverse and real-world conversational data could further enhance robustness. Additionally, the current prototype supports only text-based interactions and does not include features such as voice input, text-to-speech, or session memory. Future work should explore the integration of speech recognition, dialogue history management, and multilingual support to accommodate users who alternate between Amharic and English.

In conclusion, the proposed BiLSTM-based Amharic chatbot demonstrates that deep learning, combined with linguistic processing and expert input, can significantly advance mental health support in low-resource language settings. The results validate both the technical robustness and social impact of the system, highlighting its potential to provide scalable, stigma-free psychological assistance in Ethiopia.

Conclusions

Mental illness, particularly depression and anxiety, significantly affects many individuals, and due to the pressures of daily life, no one is entirely free from its impact. In Ethiopia, mental health care primarily depends on human professionals, but access to effective therapy is limited by a shortage of professionals, limited resources, economic constraints, fear of social stigma, and misconceptions about mental disorders. To address these barriers, we developed a retrieval-based Amharic-language chatbot using NLP and a BiLSTM architecture, aimed at providing accessible, stigma-free psychological support. The model was trained on 4,450 manually prepared entries, created with input from mental health professionals to ensure clinical relevance. Preprocessing steps included normalization, tokenization, stop-word removal, and Word2Vec embeddings. The BiLSTM model achieved 91.25% accuracy and was positively evaluated in UAT, with 86.6% satisfaction reported by experts and volunteers, demonstrating its usability and real-world potential.

This research makes a novel contribution by demonstrating how deep learning can be adapted to Amharic, a morphologically rich but low-resource language with scarce digital mental health datasets. Unlike prior English-language or rule-based systems like ELIZA and ALICE, this work introduces an advanced deep learning-based approach to a linguistically complex and culturally sensitive context, addressing both technical and societal gaps in the literature. However, the study is limited by the relatively small training dataset and the use of a retrieval-based model, meaning it can only respond to queries represented in the training dataset and may struggle with unanticipated user inputs. Future work could address these limitations by expanding the dataset, incorporating generative models for more flexible and context-aware responses, and supporting multi-turn dialogues for a more natural conversational experience. Integrating voice-based interactions would enhance accessibility, and extending the system to other Ethiopian languages such as Tigrigna, Afan Oromo, and Sidama would promote broader inclusivity and impact. Together, these advancements would make the chatbot a more robust, interactive, and widely useful tool for mental health support in Ethiopia.

Appendices

As part of our research thesis at the University of Gondar, we are conducting user acceptance testing that investigates the importance of a developed prototype. We need your assistance and appreciate your willingness to complete the questionnaire Answer this study that can be identified with you will remain confidential. answer the questions below by marking "✓".

Respondents Name Endalkachew Sisay M

Educational level -BSC [ ] [✓] PHD [ ] other [ ]

No	Evaluation criteria	pcor	fair	'gocd	Vary	cscall cıcaill	ocsclioir al	Λıvcııg .
1	Does the prototype is user frindly to end users in terms of usage! How to talk casily?						✓	
2	Does the prototype respond with clear language?						✓	
3	Does the prototype have a better response?			✓				
4	Does the prototype have a better response time?			✓				
5	How do you explain the importance of this prototype?						✓	
	Average performance							

- condition is needed After Question Type Respond of the progru

How to cite this article:
Wubneh G D, Askale G T, Woldeyohannis M M, et al. (February 25, 2026) Design and Development of Amharic Chatbot for Mental Health Using Bidirectional Long Short-Term Memory. Cureus J Comput Sci 3 : es44389-026-00034-2. DOI https://doi.org/10.7759/s44389-026-00034-2

FIGURE 5: User acceptance sample test result

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Gashaw D. Wubneh, Michael M. Woldeyohannis, Worku A. Degife

Acquisition, analysis, or interpretation of data: Gashaw D. Wubneh, Getnet T. Askale

Drafting of the manuscript: Gashaw D. Wubneh

Critical review of the manuscript for important intellectual content: Gashaw D. Wubneh, Getnet T. Askale, Michael M. Woldeyohannis, Worku A. Degife

Supervision: Michael M. Woldeyohannis

Disclosures

Human subjects: Consent was obtained or waived by all participants in this study. University of Gondar psychotherapy Team issued approval 4. The research protocol and prototype were approved by the University of Gondar Psychotherapy Team IRB, ensuring compliance with ethical standards for research involving human participants. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

First and foremost, I would like to thank the almighty GOD and Holy Mother for making all things possible, granted in my life, giving me the strength, courage, patience, and perseverance to endure this research study. I am deeply grateful for the wise counsel of my advisor, Michael Melese (Ph.D.), and Worku Abebe (Ph.D.), for their guidance, suggestions, and support throughout the preparation of this thesis. They patiently listened to the challenges I encountered and showed me how to overcome them, consistently providing constructive comments that greatly improved the quality of this study. Their critical assessments, feedback, and suggestions helped me stay on course and ensured that my research remained meaningful. In addition, I would like to express my sincere gratitude to Mr. Getnet Tigabie for his support, constructive suggestions, and encouragement. I would also like to express my deepest gratitude and indebtedness to my Brother, Endalewu Desalegn, and staff member Menbere Hailu for their unlimited support and constant encouragement for my achievements. Finally, I would like to thank all my dear friends and class reunions for the wonderful times we shared, for your genuine support, and for the joy and positivity you brought into my life. The datasets and codes generated and/or analyzed during this study are included in this article or are available from the corresponding author upon reasonable request.

References

1. Ahmed A, Hassan A, Aziz S, et al.: Chatbot features for anxiety and depression: A scoping review. Health Informatics Journal. 2023, 29:1-17. 10.1177/14604582221146719

2. Sharma H, Puri P, Rawat D: Digital psychiatry - curbing depression using therapy chatbot and depression analysis. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India. 2018, 627-631. 10.1109/ICICCT.2018.8472986

3. Godse NA, Deodhar S, Raut S, Jagdale P: Implementation of chatbot for ITSM application using IBM Watson. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India. 2018, 1-5. 10.1109/ICCUBEA.2018.8697411
4. Memon Z, Tahir F, Dehraj S, Noureen F, Jalbani AH, Bux K, JalbaniGH: Multi interactive chatbot communication framework for health care. IJCSNS International Journal of Computer Science and Network Security. 2020, 20:121-124.
5. Santoso HA, Winarsih NAS, Mulyanto E, et al.: Dinus intelligent assistance (DINA) chatbot for university admission services. 2018 International Seminar on Application for Technology of Information and Communication, Semarang, Indonesia. 2018, 417-423. 10.1109/ISEMANTIC.2018.8549797
6. Manoj Kumar V, Keerthana A, Madhumitha M, Valliammai S, Vinithasri V: Sanative chatbot for health seekers. International Journal of Advanced Trends in Computer Science and Engineering. 2016, 5:16022-16025.
7. Oh KJ, Lee D, Ko B, Choi HJ: A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. 2017 18th IEEE International Conference on Mobile Data Management (MDM), Daejeon, Korea (South). 2017, 371-375. 10.1109/MDM.2017.64
8. Thorat SA, Jadhav V: A review on implementation issues of rule-based chatbot systems. SSRN Electronic Journal. 2020, 1-6. 10.2139/ssrn.3567047
9. Weizenbaum J: ELIZA—A computer program for the study of natural language communication between man and machine. Communications of the ACM. 1966, 9:36-45. 10.1145/365153.365168
10. Abu Shawar B, Atwell E: Chatbots: are they really useful?. Journal for Language Technology and Computational Linguistics. 2007, 22:29-49. 10.21248/jlcl.22.2007.88
11. Srivastava P, Singh N: Automatized medical chatbot (Medibot). 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), Mathura, India. 2020, 351-354. 10.1109/PARC49193.2020.236624
12. Rahman MM, Amin R, Khan Liton MN, Hossain N: Disha: An implementation of machine learning based Bangla healthcare chatbot. 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh. 2019, 1-6. 10.1109/ICCIT48885.2019.9038579
13. Abu Shawar B, Atwell E: ALICE chatbot: Trials and outputs. Computación y Sistemas. 2015, 19:625-632. 10.13053/cys-19-4-2326
14. Bisandu DB: Design science research methodology in Computer Science and Information Systems. International Journal of Information Technology. 2016, 1-6.
15. Schorr F, Hvam L: Design science research: a suitable approach to scope and research IT service catalogs. 2018 IEEE World Congress on Services (SERVICES), San Francisco, CA, USA. 2018, 25-26. 10.1109/SERVICES.2018.00026
16. Australian Institute of Professional Counsellors: AIPC's Case Study Collection. Australian Institute of Professional Counsellors, Queensland, Australia; 2007.
17. Bunge EL, Mandil J, Consoli AJ, Gomar M: CBT Strategies for Anxious and Depressed Children and Adolescents: A Clinician's Toolkit. Guilford, New York, NY; 2017.
18. Raj S: Building Chatbots with Python: Using Natural Language Processing and Machine Learning. Apress, Berkeley, CA; 2019. 10.1007/978-1-4842-4096-0
19. Liu H: Sentiment analysis of citations using Word2vec. arXiv. 2017, 10.48550/arXiv.1704.00177
20. Ma W, Cui Y, Liu T, Wang D, Wang S, Hu G: Conversational word embedding for retrieval-based dialog system. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020, 1375-1380. 10.18653/v1/2020.acl-main.127
21. Adamopoulou E, Moussiades L: Chatbots: History, technology, and applications. Machine Learning with Applications. 2020, 2:100006. 10.1016/j.mlwa.2020.100006
22. Verma D, Yadhav R, Sanjana M, Varsha TP, Vineeth: A ChatBot using Seq2Seq and bag of words model. International Research Journal of Engineering and Technology. 2020, 7:1751-1754.
23. Tonja AL, Belay TD, Azime IA, Ayele AA, Mehamed MA, Kolesnikova O, Yimam SM: Natural language processing in Ethiopian languages: current state, challenges, and opportunities. arXiv. 2023, 10.48550/arXiv.2303.14406
24. Patel F, Thakore R, Nandwani I, Bharti SK: Combating depression in students using an intelligent chatbot: a cognitive behavioral therapy. 2019 IEEE 16th India Council International Conference (INDICON), Rajkot, India. 2019, 1-4. 10.1109/INDICON47234.2019.9030346

How to cite this article:
Wubneh G D, Askale G T, Woldeyohannis M M, et al. (February 25, 2026) Design and Development of Amharic Chatbot for Mental Health Using Bidirectional Long Short-Term Memory. Cureus J Comput Sci 3 : es44389-026-00034-2. DOI https://doi.org/10.7759/s44389-026-00034-2