The Indonesian Journal of Computer Science
www.ijcs.net
Volume 15, Issue 5, October 2026
https://doi.org/10.33022/ijcs.v15i5.5232

Answer-aware Question Classification in the Medical Domain

Berhanu Bogale[1*], Solomon Teferra[2], Gebeyehu Belay[3], Tesfa Tegegne[4]
berhanubogale0101@gmail.com[1]
[1,3,4]Department of Computer Science, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia
[1]Department of Information Technology, Gafat Institute of Technology, Debre Tabor University, Debre Tabor, Ethiopia
[2]Department of Computer Science, School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia

Article Information

Received : 25 Aug 2026
Revised : 27 Sep 2026
Accepted : 23 Oct 2026

Keywords

Amharic Question Answering, Deep Learning; BiLSTM, Healthcare NLP, Low-Resource Language, Question Classification

Abstract

Amharic question answering has received limited attention compared with high-resource languages due to the scarcity of annotated datasets. This study investigates Amharic question type classification using two experimental settings: Question Classification (QC) and Question–Answer Classification (QAC), to evaluate the impact of answer-aware contextual enrichment on classification performance. The experiments were conducted on the AmHQA dataset, which consists of four question categories: definition, description, list, and factoid. Two deep learning approaches were employed. The first combines pre-trained AmharicWord2Vec embeddings with a Bidirectional Long Short-Term Memory (BiLSTM) network enhanced by an attention mechanism. The second employs Bidirectional Encoder Representations from Transformers (BERT) to leverage contextualized semantic representations through self-attention. In the QC phase, only question text is used as input, whereas the QAC phase incorporates both question and answer text as a unified input sequence. Model performance was evaluated using accuracy, precision, recall, F1-score, classification reports, and confusion matrix analysis. The BERTbased QAC model achieved the highest performance, outperforming the Word2Vec-BiLSTMAttention architecture. The findings confirm that answer-aware contextual information substantially enhances Amharic question type classification.

Attribution-ShareAlike 4.0 International License

A. Introduction

The rapid advancement of Natural Language Processing (NLP) has significantly transformed the development of intelligent systems capable of understanding and responding to human language[1]. Among these applications, Question Answering (QA) systems have gained considerable attention due to their ability to provide precise and contextually relevant answers to user queries[2][3]. QA systems have been successfully deployed in diverse domains such as healthcare, education, and customer support enabling efficient access to structured and unstructured knowledge sources[4]. Medical QA systems have become important as they support healthcare professionals and patients by providing timely access to health-related information[5][6].

The healthcare sector particularly requires robust language technologies capable of processing medical information in local languages[7] [10][11]. Access to reliable medical information in Amharic is essential for improving healthcare communication, supporting patient education, and facilitating digital health services in Ethiopia[8][9].

Question classification is a fundamental component of QA systems, as it enables the system to determine the intent and type of user questions, thereby improving answer retrieval accuracy[12]. For example, if the system understands the question "የወባ በሽታ ምልክቶች ምንድን ናቸው?" (What are the symptoms of Malaria?), the classification module identifies the intent as "Symptoms". Instead of searching through the entire history of Malaria or its treatments, the system focuses strictly on clinical manifestations like fever and chills. If the question is "ፓራሴታሞል በቀን ስንት ጊዜ ይወሰዳል? (How many times a day should Paracetamol be taken?), the system classifies this under "factoid/Dosage/Administration". This helps the retriever looks for quantitative guidelines rather than general drug descriptions. When asked "ስኳር በሽታ ምንድን ነው? (What is Diabetes?)", the system recognizes the need for a "Definition". This will provide a high level educational summary instead of a list of items. However, the absence of annotated Amharic medical QA datasets has hindered research efforts toward building accurate medical question classification and answering systems.

To assess the dataset, we experimented with several deep learning models, namely LSTM, BiLSTM, and BiGRU combined with Word2Vec embeddings, and compared their performance with the transformer-based BERT model. Precision, recall, F1-score, and confusion matrices were used to measure how accurately they classified the different question categories. The results revealed that BERT delivered the highest overall performance, clearly surpassing the recurrent neural network-based models. These outcomes highlight the strength of contextualized transformer representations in handling Amharic medical question answering systems.

Our contributions can be summarized as follows :(1) we constructed a new dataset of 2690 Amharic health question–answer-category triplets for question type classification and answer selection task. (2) We conduct a detailed statistical analysis of question–answer patterns using Python. (3) We evaluate BERT and compare it with five deep learning models on the developed dataset. BERT model becomes a better algorithm in classification performance.

The rest of our work is organized as follows. Following the discussion of question answering system, Section 2 discusses the related works. The third section discusses the methodology in which the five deep learning models are described. The fourth section discusses experimental settings, and analyses of the experiment results. The 5th section introduce the conclusion of our work.

A.1 Question answering system

A typical question answering (QA) system adopts a pipeline architecture composed of three core components: question analysis, document analysis, and answer analysis (see figure 1) [13][14]. Modules are arranged in such a way that the output of one module feeds into the next.

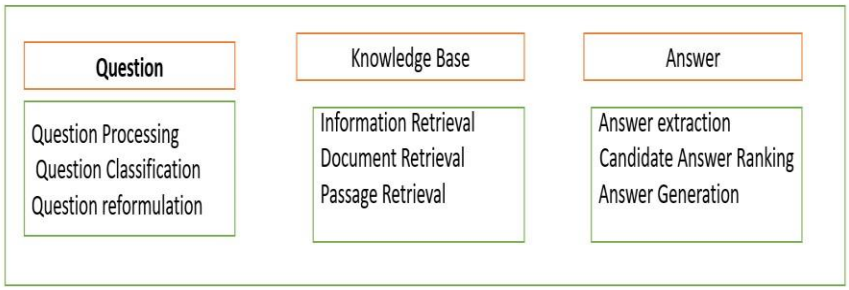

Figure 1. General architecture of question answering (src: as designed by authors)

The implementation of question classification is a computer program to identify the type of answers. Question classification suggest different processing strategies and can be used as a vital filter. The 9 types of health questions such as how, how many, who and how much type question questions. By categorizing a medical questions into specific semantic domains such as symptoms (ምልክቶች), treatments (ሕክምና), or preventive measures (መከላከያ), the system can significantly narrow the search space for plausible answers.

B. Related work

B.1 Approaches

B.1.1. Machine learning approaches

Question answering (QA) systems have become a major research area in natural language processing (NLP) due to their ability to automatically respond user queries. The introduction of machine learning (ML) marked an important shift in the field, as algorithms such as k-nearest neighbors (KNN) and support vector machines (SVMs) enabled better scalability and higher accuracy by learning patterns directly from data[13]. These methods also introduced feature selection techniques to handle challenges like high-dimensional feature spaces and improve computational efficiency. As a result, they significantly enhanced classification performance and adaptability([15][16][17]. Amartya Roy et al. [18] conducted a study on Bangla text classification using Naive Bayes (NB), Support Vector Machine (SVM), Hidden Markov Model (HMM), and Random Forest (RF). Li and Roth et al. [19] created the first of its kind question classification system using a hierarchical classifier SNoW learning architecture

for fine grained categories. They assembled the UIUC question classification dataset, which contains 5,500 training questions and 500 test questions. However, manual feature selection in machine learning limits their ability to process large amount of textual data. Their performance was less due to insufficient contextual representation (e.g[14], [20] and [21]).

B.1.2. Deep learning Approaches

More recently, deep learning has further transformed text classification by enabling models to learn complex semantic relationships within text data. Approaches such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models like BERT have substantially advanced performance by capturing deeper contextual and sequential information[22]. Long Short-Term Memory (LSTM) improves the problem of RNN gradient explosion and gradient disappearance by changing the layer structure. Liu et al. [24] used attention-based BiGRU-CNN network to classify Chinese questions. The hybrid design effectively balances context understanding (via BiGRU) and feature representation (via CNN and attention). Yilmaz et al. [25] contributed a Turkish question classification system using a hybrid of CNN-LSTM, CNN-SVM and get an accuracy of 94%. A study on Malayalam Question Answering (QA) [4] proposed a hybrid of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks using the Facebook BAbI dataset. CNN is first used to extract high-level semantic features from the text, while the Bi-LSTM layer models contextual dependencies and sequential relationships. A Soft Max activation function is applied to predict the correct answers for the corresponding questions. The experimental results demonstrate that the hybrid CNN-BiLSTM model achieves 91% accuracy across all five tasks. The study presents a valuable contribution to the advancement of question classification. Transformer-based models such as BERT have achieved state-of the art performance across various tasks such as question classification[26], machine reading[27]. The work in [28] classifies Amharic community-based questions collected from social media platforms using transfer learning approach. The followings gaps were identified in the existing works:

Language coverage: The dataset focuses only on Amharic, leaving other Ethiopian languages and health care domains. This limits cross lingual generalization and multilingual benchmarking. Most restricted to English while others focus on Chinese language, no multilingual or domain transfer. Additionally, many of the approaches still faced long range dependency limitations[29].

Domain scope: Low resource languages especially Amharic research remains limited due to the scarcity of annotated datasets and domain-specific corpora. Moreover, existing Amharic NLP studies have focused on sentiment analysis[30], text classification, machine translation, and general domain question answering[21][31]. The open domain questions are primarily drawn from limited sources (Wikipedia news articles) which may not capture clinical or community health contexts. To address this gap, we introduce Amharic medical question answering dataset and evaluates both recurrent neural

network models and transformer-based BERT model. The following tables shortens the literature reviewed paper (table 1).

Table 1: Summary of literature review in different modalities

Aspect	**AmQA Benchmark (2024)**[21]	**Chinese QA**[22]	**LSTM + Attention Text Classification (IEEE)**[21]
Language Coverage	Focused solely on Amharic, a low-resource African language.	Chinese, using Fudan University's dataset (17,252 questions).	General text datasets (not language-specific).
Task Scope	QA benchmarking for low-resource NLP.	Question classification (categorization task).	Broad text classification (sentiment, spam, etc.).
Dataset Scale	Relatively small compared to global QA datasets; limited domain diversity.	Medium-sized dataset, but single-language and single-domain.	Larger, general-purpose datasets; not specialized.
Methodology	Benchmark creation, evaluation of QA models on Amharic.	Attention-based LSTM for classification.	Hybrid CNN + LSTM + Attention for feature extraction.
Comparative Baselines	Compared against other QA systems, but lacks multilingual baselines.	Compared against traditional classifiers (SVM, CNN, and RNN).	Compared against CNN, RNN, and plain LSTM.
Limitations/Gaps	No multilingual extension, limited clinical/real-world validation.	Restricted to Chinese, no multilingual or domain transfer.	No benchmarking against Transformerbased models (e.g., BERT).

C. Methods

This study evaluates Amharic question type classification using two experimental phases, namely Question Classification (QC) and Question–Answer Classification (QAC). In the QC phase, only the question text is used as the input feature, where each question is mapped directly to one of the four predefined AmHQA categories: definition, description, list, and factoid. In the QAC phase, the question and its corresponding answer are concatenated into a single input sequence to provide additional semantic context for classification. Two classification approaches are employed in both phases. The first approach uses AmharicWord2Vec embeddings combined with a Bidirectional Long Short-Term Memory (BiLSTM) network enhanced by an attention mechanism, which captures sequential contextual dependencies while emphasizing informative tokens. The second approach utilizes a BERT-based transformer model, which leverages contextualized deep semantic representations through self-attention. Each model evaluated using accuracy, precision, recall, F1-score, classification report, and confusion matrix analysis. This dual-phase evaluation enables

comparative analysis of whether answer-aware contextual enrichment improves question type classification performance in low-resource languages like Amharic.

Two Approaches (Two Phases)

Type 1: QC (Question Classification)

Input: Question → Category

Type 2: QAC (Question–Answer Classification)

Input: Question + Answer → Category

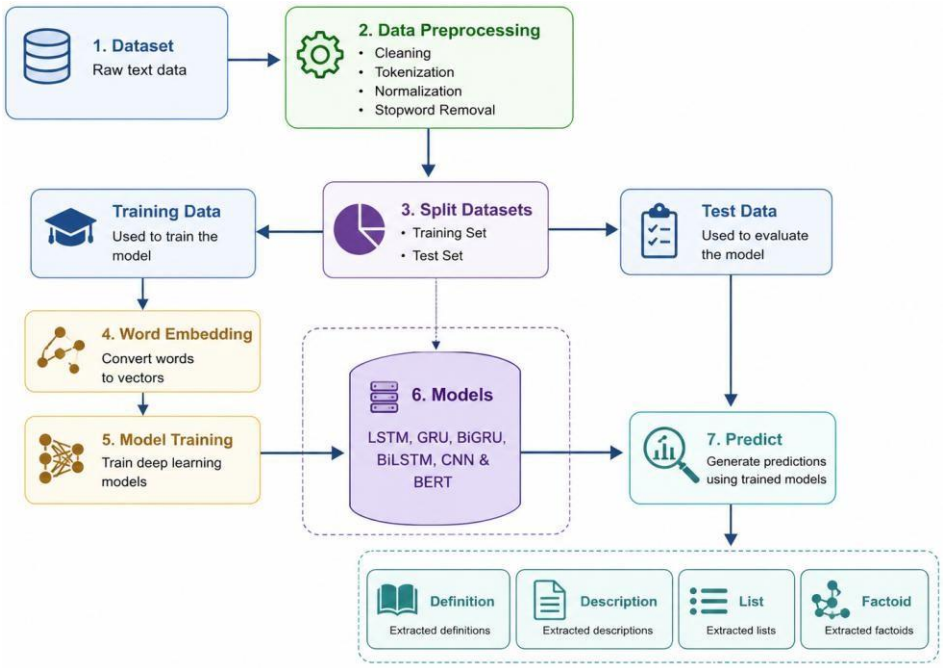

Figure 2. Proposed architecture for Amharic question type classification

The traditional deep learning models are structured as follows:

Input Layer: The input layer of traditional deep neural network contains the word sequences in a question Q. We represent each word in the question as a fixed-size vector from pre-trained Word2Vec embeddings. The dimensions of the word vectors and hidden layers of LSTMs are set as 100. The initial word embeddings are trained on both the unlabeled and labeled reviews using word2vec.

LSTM/BiLSTM/BiGRU Layer: For this study, we use four deep learning models to model the input sequences. In BiLSTM/BiGRU (see in figure 2), there are two layers of hidden nodes from two directions to capture different sequences. For instance, of the two LSTMs, the first hidden layers have recurrent connections from the past words. The second LSTM takes in the other directions. The forward LSTM network take forward hidden state $\overrightarrow{h}_{i,j}$ from the forward

LSTM network and the backward hidden state $\overleftrightarrow{h}_{i,j}$ from the backward LSTM network. We represent the final state at by concatenating the forward hidden state and backward hidden state.

Thus $h_{i,j} = \overrightarrow{h}_{i,j} || \overleftrightarrow{h}_{i,j}$.

BERT Architecture for classification

A pre-trained BERT (Bidirectional Encoder Representations from Transformers) [8][9] is based on a multi-layer bidirectional transformer, trained on clean plain text. BERT understand the meaning of a word within a given sentence. BERT architecture contains twelve transformer blocks and twelve self-attention heads with a hidden state of 768. For classification of text, it represents an entire sequence by the final hidden state h of its token [CLS]. Finally, a softmax classifier is used for predicting the category.

C.1 Dataset

The Amharic health QA (AmHQA) dataset was developed to perform the question type classification and answer selection tasks. The dataset consists of 2690 question answer pairs labeled by two annotators. We prepare the dataset from Amharic Wikipedia, FanaBroadcasting, Ameco, and BBC Amharic website posts using 2 Amharic-speaker crowd workers. This Amharic health QA dataset is original and human-annotated by two individual annotators, with a pairwise agreement of 74.9%. We partitioned the dataset with a ratio of 90:10 for training, and testing respectively.

The system first predicts the question category are labeled as factoid, definition, description and list.

a. Factoid questions: It is a simple question that requires a single short phrase or sentence to be answered. It is expected to answer entity related questions such as what/የት/, When/መቼ/, how many/ስንት/, who/ማነው/ question types. Non-factoid questions ("/በኢትዮጵያ ያለው የኤች አይቪ መንስኤ ምንድነው?/What are the causes of HIV in Ethiopia?") are typically considered much more complex than factoid questions.
b. Descriptive questions: descriptive questions often contains “why/ለምን/” and “how/እንዴት/” question types. The typical responses to this type of question include reasons, explanations, and elaborations.
c. List type questions: For list-type questions, the answers are expected as a list of facts or entities, e.g. List of diseases known in 2021/በ2021 የታወቁ በሽታዎች/.
d. Definition type questions: The question “የጨጓራ በሽታ ህክምናው ምንድን ነው?/What is the treatment for stomach disease?”), the word “ምንድን ነው” can determine the expected answer type as a treatment-related definition.

C.1.1. Data Preprocessing

The AmHQA dataset is preprocessed through text cleaning, tokenization, label encoding, and train-test splitting. The data preprocessing includes mainly the following:

Cleaning: cleaning the data involves eliminating any English letters ([A-Za-z]) and Arabic script characters to retain only relevant Amharic text. Special characters that do not fall within the Amharic Unicode range ([^ፀ-? 0-9\s]) are eliminated. Rows containing empty values in crucial columns were identified

and removed using a predefined function. This function detect missing values and eliminate the corresponding rows. To ensure traceability, the original dataset was preserved under a new filename before saving the cleaned version as a CSV file. Duplicate question–answer pairs are also removed to avoid redundancy.

Normalization: the Amharic character normalization was made due to consistent writing styles like አለም፣ ዐለም and ዓለም, substitute {አ፣ ዐ፣ዓ} with አ. Character normalization[32] is necessary to avoid any differences in similar words written with different characters that sound the same but have different appearances. This type of normalization first tokenizes the given text into words, then it splits each word into its characters and inspects if the character is interchangeably usable.

Tokenization: Following data cleaning, tokenization is performed on question and answer pairs. Tokenization is the process of dividing the entire corpus data into small chunks of words or sub words[33]. Both question and context are sentence-tokenized and concatenated as input to the model.

We utilize the Word Piece[34] due to its simplicity and compatibility with various PLMs[35]. For the Word2Vec-based model, tokenized sequences are padded to a fixed length and initialized using pre-trained Amharic word embeddings. For BERT, input sequences are tokenized using the BERT tokenizer with truncation and padding.

Padding: Padding is a key preprocessing step in natural language processing (NLP) that ensures all input sequences have a uniform length for efficient batch processing in deep learning models[36]. Since sentences naturally vary in length, padding adds special tokens (usually zeros) to shorter sequences so they match the longest sequence in the batch.

Data Balancing: The dataset includes three columns: question, answer, and category. The category column has a value of definition, description, factoid and list question classes. The distribution of the dataset per class is 289 are labeled as factoid, 808 as list, 463 as definition and 1131 as description. This imbalance in question distribution poses a potential bias in model training[37]. The figure 3 illustrates the unbalanced dataset with a noticeable variation. Description questions are heavily overrepresented, whereas factoid and definition questions are underrepresented. Such imbalance can become biased toward the majority class during machine learning model training [37]. Therefore, handling this imbalance using resampling methods like oversampling, under sampling, or advanced techniques such as SMOTE[38] become important.

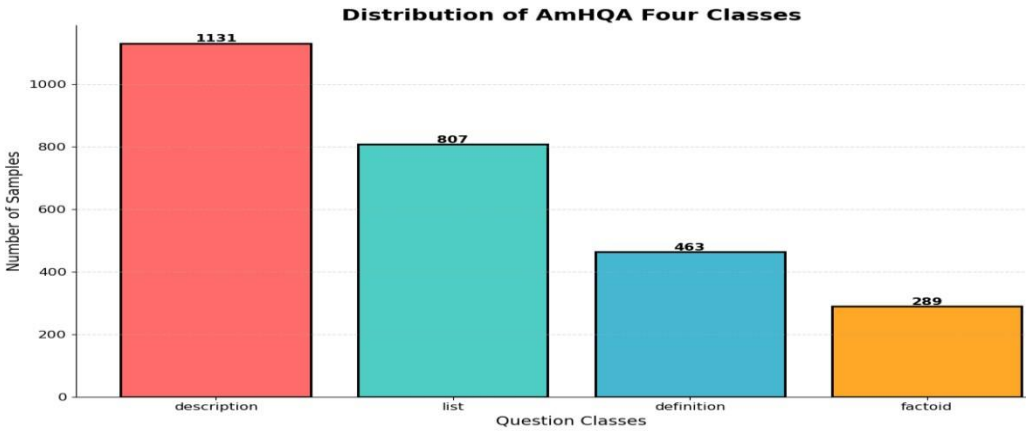

Figure 3. Distribution of questions in different in different class labels

Figure 4 illustrates the details of balanced datasets used for question classification tasks. The SMOTE (Synthetic Minority Over-sampling Technique) approaches are employed for data balancing. Unlike simple duplication, SMOTE[38] generates synthetic samples for minority classes by interpolating between existing data points. Once the dataset is balanced, the dataset was divided into training, and test sets while maintaining the original question distribution. 10% of the data was allocated to the test set, while the remaining portion was further split into training (90%). The final datasets were saved as separate CSV files—train_set.csv, and test_set.csv for subsequent modeling phases. The dataset distributions included 269 samples in the test set and 2421 in the training set, all equally distributed among the four question labels.

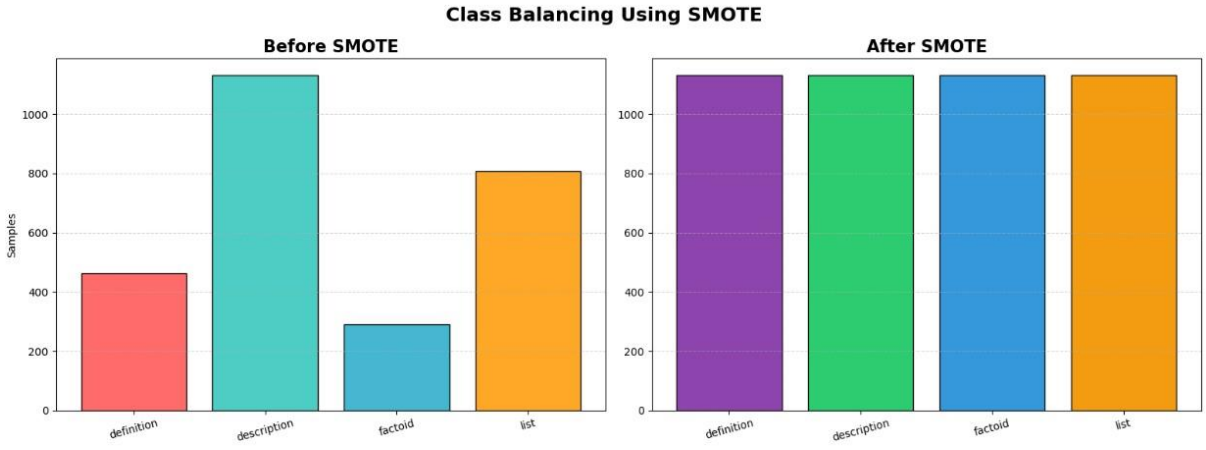

Figure 4. Class Distributions after SMOTE technique

C.2 Feature Representation

The question words are first mapped into vector representations using a pre trained embedding layer such as Word2Vec.

C.2.1. Word2Vec.

There are different methods for feature construction Word2vec, GloVe and ELMO that transform text classification[39]. These create a good semantics vocabularies for classifier learning. Word2Vec is a neural embedding model that learns continuous vector representations of words by exploiting their contextual co-occurrence patterns in large text corpora. We trained a Word2Vec model with an embedding dimension of 300 to capture rich semantic and contextual information. The training was conducted on a large Amharic text corpus, resulting in a vocabulary of 94,734 unique words. The pre-trained embedding can be used as a pre-trained representations for tasks such as question classification together with BiLSTM and CNN-based models.

C.3. Other Models

To prove the superiority of the BERT model, we compared our model with five baseline models. By comparing the performance of these baseline approaches with a transformer-based model, this study aims to highlight the trade-offs between traditional recurrent architectures and modern contextualized language models. These baselines are traditional deep learning approaches that combine static word embeddings with sequential neural networks. The BERT model was compared against five baseline models: Convolutional Neural Network (CNN)[40], Bidirectional GRU (Bi-GRU), long short-term memory (LSTM), and BiLSTM [41] models.

CNN: CNNs are good at recognizing patterns like key phrases and RNN are good to capture word orders. Recurrent Neural Networks (RNNs) are particularly effective for NLP tasks such as part-of-speech (POS) tagging and question answering, where understanding long-range dependencies and contextual relationships within text is essential. Because of their efficiency in detecting local features, CNNs have become one of the most widely adopted architectures for text classification (TC)[30] tasks.

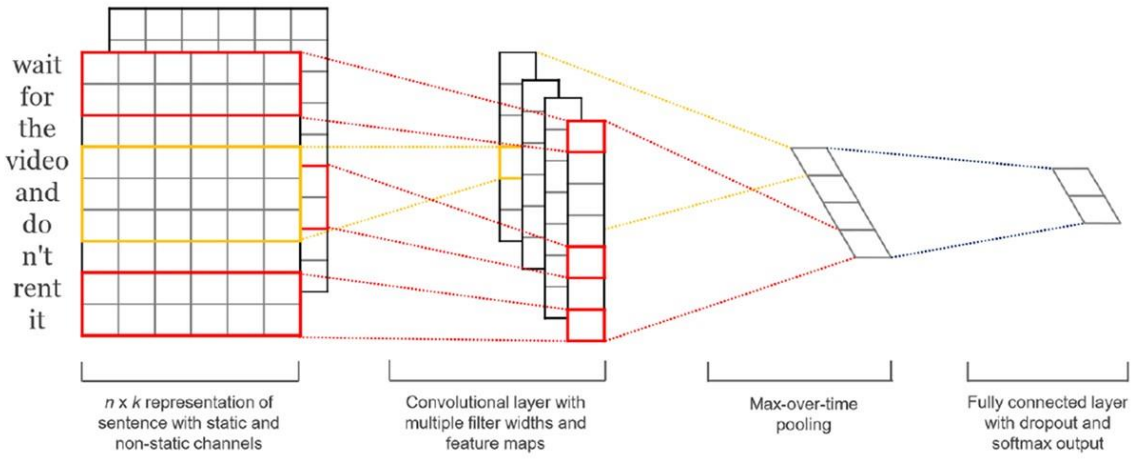

Figure 5. Architecture of a Sample CNN Model for Text Classification[30]

C.3.1. LSTM Model

The LSTM model captures sequential dependencies using memory cells and gating mechanisms. A LSTM network as shown in figure 2 contains forget gate to discard the unnecessary information from previous time steps, an input gate to select useful information from inputs, and an output gate to control the outputs of the current LSTM network. After processing all tokens, the final hidden state (h_final) represents the entire sentence meaning. The hidden vector h (t) at the time step t is updated as:

$$i_t = \delta(W_i x(t) + U_i h(t-1) + b_i) \quad (1)$$
$$f_t = \delta(W_f x(t) + U_f h(t-1) + b_f) \quad (2)$$
$$o_t = \delta(W_o x(t) + U_o h(t-1) + b_o) \quad (3)$$
$$C_t = tanh(W_c x(t) + U_c h(t-1) + b_c) \quad (4)$$
$$C_t = i_i * C_t + f_t * C_{t-1}) \quad (5)$$
$$h_t = o_t * \tanh(C_t) \quad (6)$$

Where i is input, f is forget and o is output , and c is a cell memory vector c. δ is the sigmoid function.

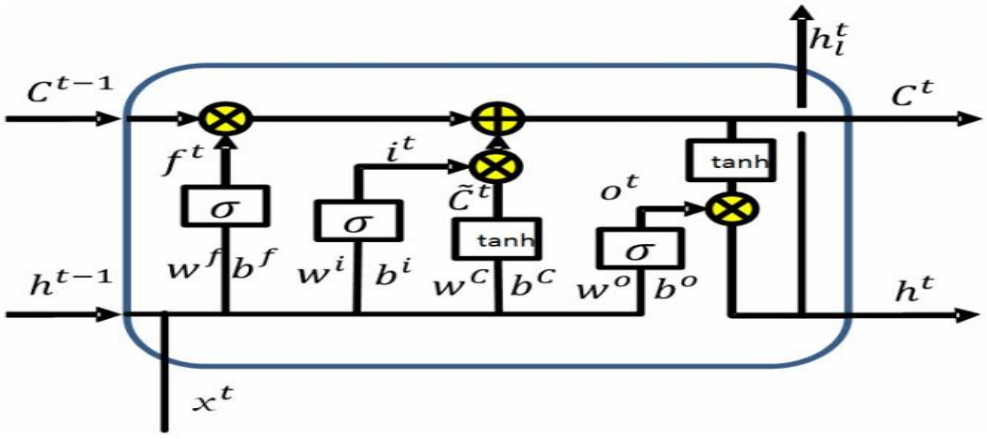

Figure 6. Structure of the LSTM[36]

C.3.2. BiLSTM model

The BiLSTM architecture is an advanced variant of the recurrent neural network (RNN) designed to effectively capture sequential dependencies in data. Unlike LSTM ,Bi-LSTM networks utilize both past and future contextual information by propagating data in forward and backward directions simultaneously[42]. The outputs from both directions are then combined, often through summation or concatenation, to enhance representational capacity. The representation of each word $h_{i(,w_j)}$ is obtained by concatenating the forward and backward hidden states as:

$$h_{i,j}^{(w)} = [\overrightarrow{BiLSTM}\left(w_{i,j}, \vec{h}_{i,j-1}^{(w)}\right); \overleftarrow{BiLSTM}(w_{i,j}, \overleftarrow{h}_{i,j+1}^{(w)})] \quad (7)$$

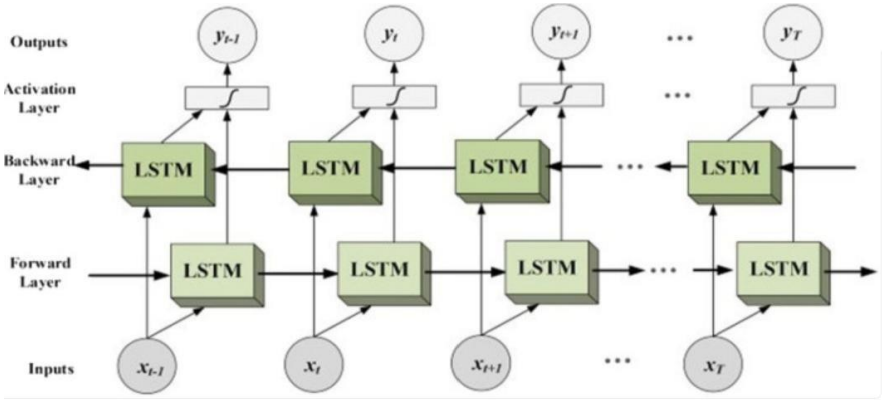

Figure 7. BiLSTM architecture

CNN+BiLSTM: The other baseline architecture is CNN+BiLSTM model as shown in Figure 8. The structure of the hybrid model is based on CNN and BiLSTM model. The CNN+BiLSTM architecture is a hybrid deep learning model that combines the strengths of Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks to improve question type classification performance. By combining CNN's ability to capture local features with BiLSTM's strength in modeling sequential dependencies, the architecture becomes more robust for complex natural language tasks. The learned representations are passed through fully connected dense layers with a softmax activation function to produce the final class predictions. This hybrid structure is particularly effective for Amharic Question Answering tasks because it captures both local phrase-level patterns and global sentence-level context.

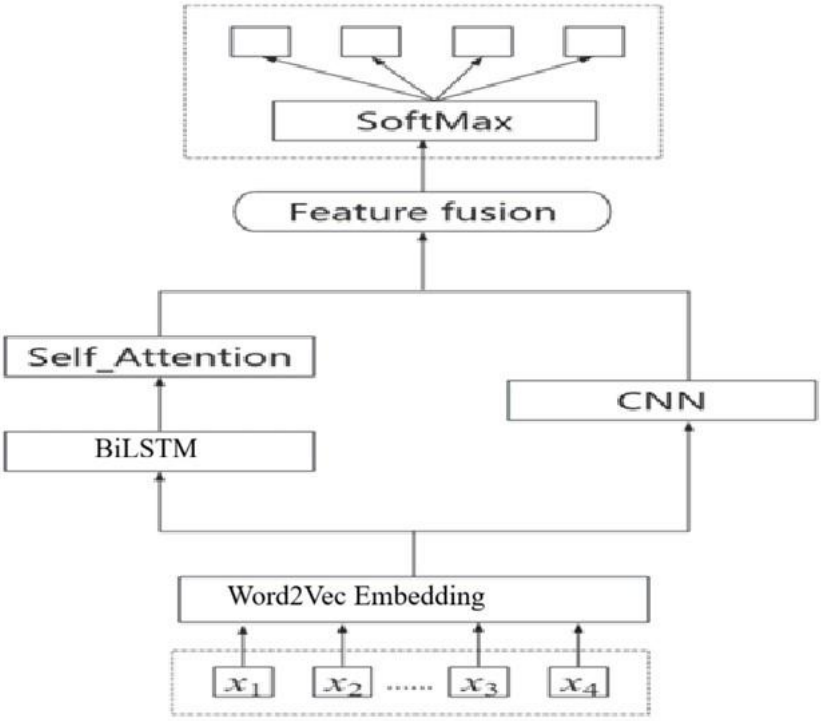

Figure 8. CNN+BiLSTM architecture (src: as authors designed)

These integrate LSTM and CNN architectures to take local and global features of sentences and documents.

D. Implementation details

To train, and test the models, we use python packages including Pytorch[43], Hugging Face[44] , Scikit-learn libraries and GPU computer. The training hyper parameters including Adam[45] optimizer, a batch of 8 and a learning rate of $2x10^{-5}$, dropout[46] of 0.3 is applied to the input of each network layer. These hyper parameters (see table 2) were selected to balance convergence and generalization.

Table 2. Possible Hyper parameters for $BERT_{base}$

Hyper parameter	Values	Description
Pre-trained Model	bert-base-multilingual-cased	Multilingual BERT model suitable for Amharic text
Task Type	Sequence Classification	Used for scoring question–answer pairs
Max Sequence Length	256	Maximum tokens for (question + candidate answer)
Tokenizer	Word Piece Tokenizer	Standard BERT sub word tokenization
Padding Strategy	Dynamic Padding	Pads input to max length per batch
Truncation	True	Cuts long sequences beyond max length

Table 3. Experimental environmental setup

Experimental environment	Details
Os	Window 11
Transformers	Transformer 4.57
Numpy	2.1.3
RAM	16G
Language	Python
Pytorch	2.8

D.2. Evaluation metrics

To evaluate the effectiveness of the selected models, a series of experiments were performed using the constructed dataset. Each model was trained using the same training and test splits for a fair comparison. The input for the models include question-category pairs adjusted based on the task requirements. The question classification architecture is composed of respective embedding layers as an input layer with dense layers and an output layer. Performance was evaluated using standard metrics like accuracy, precision, recall and F1-score.

Let TP, FP, FN, and TN represent true positive, false positive, false negative and true negative respectively. The classification Accuracy are defined as in equation 1.

$$Accuracy = \frac{(TP+TN)}{N} \quad (8)$$

Here, N represent the number of samples.

$$Accuracy = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, F1\ score = \frac{2\ xPrecision\ xRecall}{Precision+Recall} \quad (9)$$

D. Result and Discussion

D.1 Question classification Result and discussion

In section comparison experiments were conducted with the other five deep learning models with the pre-trained word2ve embeddings. Table 4 presents the performance comparison of the proposed and baseline models on the test set using accuracy, precision, recall, and F1-score.

D.1.1. Type 1: Question Classification (QC)

In the QC phase, only the question text is used as input for classification.

Table 4: QC performance Results

Model	Accuracy (%)	Precision	Recall	F1-Score
Word2Vec + LSTM + Attention	84.63	0.85	0.85	0.85
Word2Vec + GRU + Attention	85.91	0.86	0.86	0.86
Word2Vec + CNN + Attention	86.42	0.86	0.86	0.86
Word2Vec + BiGRU + Attention	87.54	0.88	0.88	0.88
Word2Vec + BiLSTM + Attention	88.42	0.88	0.88	0.88
BERT	**92.76**	**0.93**	**0.93**	**0.93**

Discussion of QC type

The results obtained in the Question Classification (QC) phase reveal that classification performance varies significantly across the evaluated architectures when only question text is used as input. Among the Word2vec-based models, BiLSTM+Attention achieved the highest accuracy (88.42%), outperforming LSTM, GRU, CNN, and BiGRU variants due to its bidirectional sequence modeling capability, which captures both preceding succeeding contextual dependencies in Amharic question sequences. The attention mechanism focuses semantically informative words that indicate question intent. The BiGRU+Attention model delivered competitive performance (87.54%). This demonstrate that bidirectional gated architectures can effectively model sequential dependencies with fewer parameters than BiLSTM. The GRU+Attention model slightly outperformed the standard LSTM+Attention model due to its simpler gating structure. The LSTM model outperformed efficient learning with limited training data. The CNN+Attention model achieved moderate performance (86.42%). CNN model shows effectiveness in capturing local n-gram patterns commonly found in Amharic interrogative structures.

However, its inability to capture long-range sequential dependencies constrained its overall classification capability. The LSTM+Attention model recorded the lowest recurrent-based performance (84.63%), highlighting the limitations of unidirectional sequence modeling for question understanding tasks that require bidirectional contextual awareness. The BERT model achieved the best QC performance (92.76%). BERT outperformed all Word2Vec-based architectures due to its transformer-based contextualized embedding mechanism.

D.1.2. Type 2: Question-Answer Classification (QAC)

In the QAC phase, both question and answer texts are concatenated and used as input.

Table 5. Type 2 Experiment results

Model	Accuracy (%)	Precision	Recall	F1-Score
Word2Vec + LSTM + Attention	88.74	0.89	0.89	0.89
Word2Vec + GRU + Attention	89.91	0.90	0.90	0.90
Word2Vec + CNN + Attention	90.42	0.90	0.90	0.90
Word2Vec + BiGRU + Attention	91.08	0.91	0.91	0.91
Word2Vec + BiLSTM + Attention	91.35	0.91	0.91	0.91
BERT	95.81	0.96	0.96	0.96

The question-answer classification (QAC) phase demonstrated improved performance across all evaluated architectures, confirming that incorporating answer information provides valuable semantic enrichment for Amharic question type classification. By combining questions with their corresponding answers, the models were able to capture richer contextual information that reduced ambiguity among classes that were difficult to distinguish in the QC phase. Among the Word2vec-based models, BiLSTM+Attention achieved the highest performance (91.35%). This shows its ability to model bidirectional sequential dependencies and identify discriminative semantic features through attention. The BiGRU+Attention model followed closely with 91.08%, indicating that GRU-based bidirectional architectures offer comparable effectiveness with lower computational complexity. The CNN+Attention model showed substantial improvement (90.42%). The answer inclusion introduces local semantic patterns that convolutional filters can effectively capture. Its inability to preserve long-range sequential dependencies limited its performance relative to recurrent bidirectional models.

GRU+Attention and LSTM+Attention exhibited notable gains, demonstrating that answer context compensates for the contextual limitations of unidirectional recurrent architectures. The BERT model achieved the best overall performance (95.91%), significantly outperforming all Word2Vec based approaches due to its transformer-based self-attention mechanism, which enables deep semantic interaction modeling between question and answer tokens. This superior performance highlights the effectiveness of contextualized transformer representations for lowresource Amharic classification tasks. Compared to QC,

the QAC phase substantially reduced confusion between definition and description categories.

D.1.3. Comparative discussion between QC and QAC

The comparison between QC and QAC reveals a consistent performance improvement across all architectures when answer information is incorporated.

Table 6: Performance Gain from QC to QAC

Model	QC Accuracy	QAC Accuracy	Improvement
LSTM + Attention	84.63	88.74	+4.11
GRU + Attention	85.91	89.91	+4.00
CNN + Attention	86.42	90.42	+4.00
BiGRU + Attention	87.54	91.08	+3.54
BiLSTM + Attention	88.42	91.35	+2.93
BERT	92.76	95.81	+3.05

The experimental findings demonstrate that incorporating answer context consistently improves classification performance across all models. This indicate that additional semantic information from answers enhances the ability to distinguish between question types. Bidirectional architectures such as BiLSTM and BiGRU consistently outperform their unidirectional counterparts (LSTM and GRU). These emphase the importance of capturing contextual information from both past and future token dependencies in Amharic question sequences. Transformer based BERT contextual embeddings become superior to all Word2Vec-based deep learning models due to their ability to model deep bidirectional semantic relationships via selfattention mechanisms. Question answer classification (QAC) setting is effective than question classification (QC) setting for Amharic question type classification. This is because the answer column provide contextual signals that enhance model discrimination. BERT models tend to perform better in our AmHQA dataset, which suggest s that the integration of questions and answers is different from that question only models. Table 6 illustrates a comparison of the results obtained by the BERT and Word2Vec versions of the QA models with higher average performance. Trained on the raw dataset and a corrected case dataset Word2Vec performed slightly less than BERT model on Amharic question dataset.

D.1.4. Class-wise Performance Analysis for the selected models

Since we have multi-class classification problem, we analyze precision and recall for each class label using three RNN models. We tabulate the performance of each model based on class labels to emphasize how the models are effective. Table 6 reports the precision, recall, f1 score and accuracies of the each model for each question category

Table 7. Classification report for 8 models with Word2Vec embeddings

Model	Item	Precision	Recall	F1-Score	Accuracy
LSTM	Definition	0.84	0.70	0.77	0.89

	Description	0.87	0.90	0.89	
	Factoid	0.93	0.98	0.95	
	List	0.92	0.93	0.92	
BiLSTM	Definition	0.87	0.92	0.83	0.92
	Description	0.85	0.91	0.88	
	Factoid	0.98	0.95	0.97	
	List	0.96	0.92	0.94	
BiGRU	Definition	0.89	0.70	0.78	0.91
	Description	0.86	0.95	0.90	
	Factoid	0.97	0.97	0.97	
	List	0.96	0.93	0.94	

AUC curve of the four model before and after data balancing in figure 11.

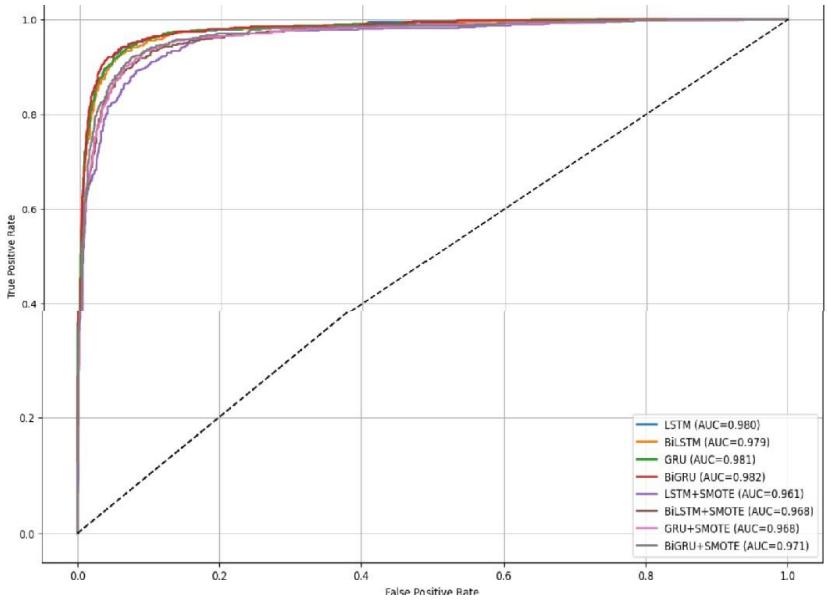

Figure 9. AUC curve of the eight models (before and after SMOTE)

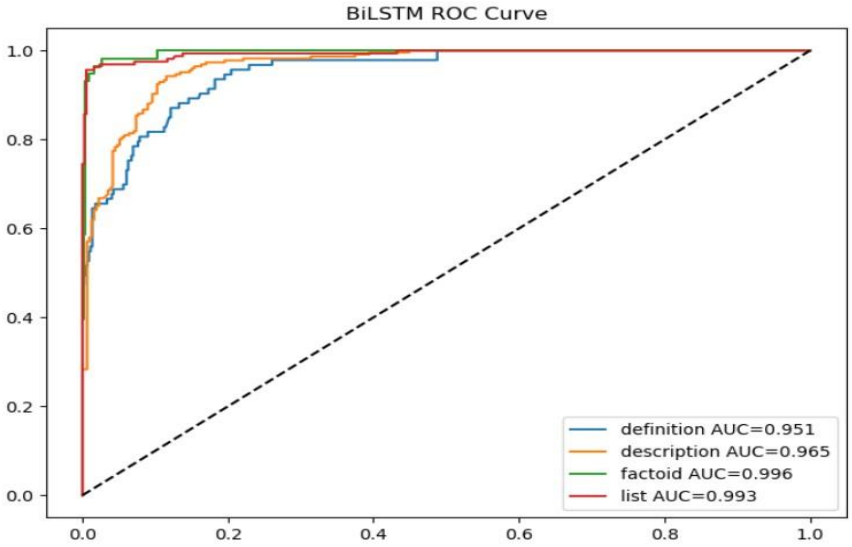

Figure 10. BiGRU after SMOTE on class level

The ROC-AUC (Figure 11) results of the BiLSTM model demonstrate excellent classification performance across all four question categories. The model achieved an AUC of 0.951 for the definition class, 0.965 for the description class, 0.996 for the factoid class, and 0.993 for the list class. These results indicate that the model has a very strong ability to distinguish between different question types, with near-perfect separability for factoid and list categories. The slightly lower performance on the definition class suggests that it is the most challenging category, likely due to its semantic overlap with description-type questions in Amharic. Overall, the consistently high AUC values (all above 0.95) confirm that the BiLSTM model has learned highly discriminative representations of the dataset and performs reliably for Amharic question classification tasks.

Figure 12 show the training and validation loss curves, respectively using BERT model. It exhibited faster convergence, achieving high performance within the first 10 epochs, while maintaining stable accuracy.

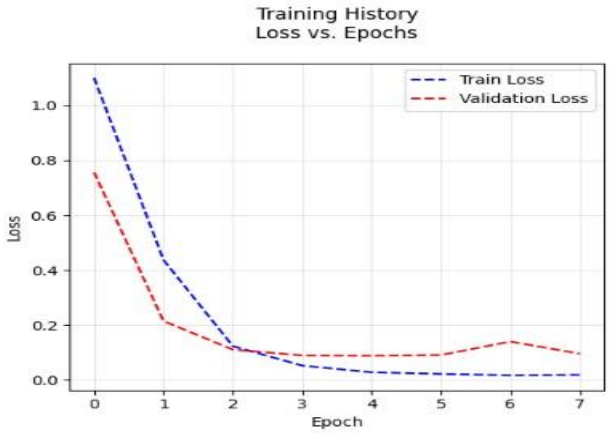

Figure 11. Training and validation Loss Curve of BERT Model

Figure 13 illustrates Confusion matrix using BERT model on Amharic medical dataset. Their error rates are similar to BiLSTM model, only the TA model shows a 1% performance increase on the question type classification capability.

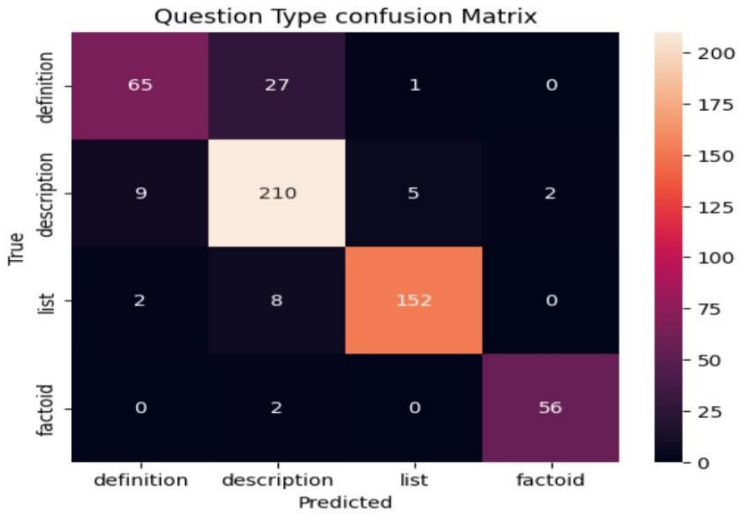

Figure 12. Confusion Matrix using BERT

The confusion matrix indicates strong overall performance of the classification model, with the highest correct predictions observed in the *description* class (900 correct) and *list* class (564 correct), showing that these categories are learned more reliably. The *factoid* class achieves moderate performance (212 correct) but is notably confused with *description* (72 cases), suggesting overlap in linguistic patterns between these two classes. The *definition* class shows comparatively weaker performance, with a substantial number of instances misclassified as *description* (138 cases), indicating difficulty in distinguishing definitional questions from descriptive ones. While the diagonal dominance confirms good classification ability, the off-diagonal errors especially between definition and description, and factoid and description highlight areas where feature similarity between classes affects discrimination.

D.1.5. Prediction Result

After the training is finished, the parameters and learned weights are saved into a file to be loaded and used to predict classes of unlabeled new data. We take the trained model as input and discover different patterns that used to predict definition, description, and factoid and list question types. Predicted answers and Labels for the selected questions using BERT model shown table 8.

Table 8. Sample Prediction Results Using BERT Model from Randomly Selected Questions

Question	Gold Answer	Predicted Label	True Label	Predicted Answer	EM	F1

በሰውነታችን ዉስጥ እረጅሙ እና ጠንካራው አጥንት ማን ነው?	የጭን አጥንት በሰውነታችን ዉስጥ እረጅሙ እና ጠንካራው አጥንት ነው	description	description	የጭን አጥንት በሰውነታችን ዉስጥ እረጅሙ እና ጠንካራው አጥንት ነው	1	0.888889	
የደም ማነስ መንስኤ ምንድን ነዉ?	በሃገራችን ዋነኛው የደም ማነስ መንስኤ ከብረት (iron) እጥረት የሚመጣ ሲሆን ይህም የተመጣጠነ ምግብ አቅርቦት በቂ ባለመሆኑና በሆድ ውስጥ ደምን የሚመጡ ትላትሎች በብዙ የሃገራችን ማህበረሰብ ውስጥ በመኖራቸው ነው።	description	description	በሃገራችን ዋነኛው የደም ማነስ መንስኤ ከብረት (iron) እጥረት የሚመጣ ሲሆን ይህም የተመጣጠነ ምግብ አቅርቦት በቂ ባለመሆኑና በሆድ ውስጥ ደምን የሚመጡ ትላትሎች በብዙ የሃገራችን ማህበረሰብ ውስጥ በመኖራቸው ነው።	1	0.962963	
ወደ አንጎል የሚደርስ አየር ማጠር አጋላጮች ምንድን ናቸዉ?	ከባድ የአስም በሽታ፣የልብ ችግር ካለ፣በጣም ዝቅተኛ የደም ግፊት መኖር፣የሳንባ በሽታ መኖር	description	description	ከባድ የአስም በሽታ፣የልብ ችግር ካለ፣በጣም ዝቅተኛ የደም ግፊት መኖር፣የሳንባ በሽታ መኖር	1		1
የፀጉር መሳሳት ምን ያመለክታል?	በጊዜ ሂደት የሚፈጠር የፀጉር መሳሳት ከጭንቀት፣ ከተመጣጠነ ምግብ እጥረት፣ አልፎ አልፎም በእርግዝና ወቅት፣ በሚወሰድ የወሊድ መከላከያ መድሃኒት፣ የክብደት መቀነስ እና መጠነኛ የጤና እክል መኖሩን ማሳያ ነው።	description	description	በጊዜ ሂደት የሚፈጠር የፀጉር መሳሳት ከጭንቀት፣ ከተመጣጠነ ምግብ እጥረት፣ አልፎ አልፎም በእርግዝና ወቅት፣ በሚወሰድ የወሊድ መከላከያ መድሃኒት፣ የክብደት መቀነስ እና መጠነኛ የጤና እክል መኖሩን ማሳያ ነው።	1		1
የሬቢስ በሽታ መተላለፊያ መንገዱ ምንድን ነዉ?	በቫይረሱ ከተጠቁ እንስሳት በተለይ ከውሾች ወደ የሰው ልጅ በንክሻ ወይም በመቧጨር ከአፋቸው በሚወጣ ፈሳሽ (saliva) በቀጥታ ቆዳን ዘልቆ የሚገባ ከሆነ ወይም በሰወነታችን ትኩስ ቁስል (fresh skin wounds) ካለ ሊተላለፍ ይችላል።	description	description	በቫይረሱ ከተጠቁ እንስሳት በተለይ ከውሾች ወደ የሰው ልጅ በንክሻ ወይም በመቧጨር ከአፋቸው በሚወጣ ፈሳሽ (saliva) በቀጥታ ቆዳን ዘልቆ የሚገባ ከሆነ ወይም በሰወነታችን ትኩስ ቁስል (fresh skin wounds) ካለ ሊተላለፍ ይችላል።	1	0.966667	
የጃርዲያ ምርመራዎች ምንድን ናቸዉ?	አብዛኛውን ጊዜ ግን የሠገራ ምርመራ በማድረግ በአይነምድር ውስጥ ተባዪን በማየት ወይም በማግኘት ሲሆን።	description	description	አብዛኛውን ጊዜ ግን የሠገራ ምርመራ በማድረግ በአይነምድር ውስጥ ተባዪን በማየት ወይም በማግኘት ሲሆን።	1		1

ዳክታል ካርሲኖማ ማለት ምን ማለት ነዉ?	ዳክታል ካርሲኖማ-ይህ የካንሰር አይነት ፈሳሽ ሊያስተላልፍ የሚችለውን የጡት አካል የሚያጠቃ ሲሆን ወደ ውጪኛው ክፍል መጥቶ የመባዛት ሁኔታው ዝቅተኛ ነው። ይህ አይነቱ የጡት ካንሰር እንደ ኢንቫሲይቭ ዳክታል ካርሲኖማ እና እንደ ኔፖል ፓጌት ዲዝዝ በብዛት የሚከሰት አይደለም። ከአስር የወንዶች ጡት ካንሰር ህሙማን መካከል አንዱ ብቻ በዚህ አይነቱ የጡት ካንሰር ህመም ተጠቂ ይሆናል። ይህ የካንሰር አይነት በቀዶ ጥገና የመዳን እድሉ ከፍተኛ ነው።	definition	definition	ዳክታል ካርሲኖማ-ይህ የካንሰር አይነት ፈሳሽ ሊያስተላልፍ የሚችለውን የጡት አካል የሚያጠቃ ሲሆን ወደ ውጪኛው ክፍል መጥቶ የመባዛት ሁኔታው ዝቅተኛ ነው። ይህ አይነቱ የጡት ካንሰር እንደ ኢንቫሲይቭ ዳክታል ካርሲኖማ እና እንደ ኔፖል ፓጌት ዲዝዝ በብዛት የሚከሰት አይደለም። ከአስር የወንዶች ጡት ካንሰር ህሙማን መካከል አንዱ ብቻ በዚህ አይነቱ የጡት ካንሰር ህመም ተጠቂ ይሆናል። ይህ የካንሰር አይነት በቀዶ ጥገና የመዳን እድሉ ከፍተኛ ነው።	1	0.813559
የእግር ጉዞ ማድረግ ጥቅም አለው?	በሳምንት ለሶስት እና አራት ሰዓት ያህል ፈጣን የእግር ጉዞ ማድረግ በልብ ህመም የመያዝ ዕድልን በ65% ይቀንሳል።	description	description	በሳምንት ለሶስት እና አራት ሰዓት ያህል ፈጣን የእግር ጉዞ ማድረግ በልብ ህመም የመያዝ ዕድልን በ65% ይቀንሳል።	1	1
ከቆዳ ሰር ያለ እብጠት ማለት ምን ማለት ነዉ?	ከቆዳ ሰር ያለ እብጠት (LIPOMA) ይህ እብጠት በቆዳ እና በጡንቻ መካከል የሚክስት የስብ ክምችት ሲሆን ሲነካ የድፎ ዳቦ አይነት ይዘት አለው ሁኖም ገነ ሲነካ ህመም አልባ ከመሆኑ ባሻገር በጣቶቻችን ስንጫነው በቀላሉ ይንቀሳቀሳል።	definition	definition	ከቆዳ ሰር ያለ እብጠት (LIPOMA) ይህ እብጠት በቆዳ እና በጡንቻ መካከል የሚክስት የስብ ክምችት ሲሆን ሲነካ የድፎ ዳቦ አይነት ይዘት አለው ሁኖም ገነ ሲነካ ህመም አልባ ከመሆኑ ባሻገር በጣቶቻችን ስንጫነው በቀላሉ ይንቀሳቀሳል።	1	0.9375
ለልጆች የጸሃይ ብርሃን እጥረት መንስኤ ምንምን ነዉ?	ቫይታሚን ዲ ያነሰው ምግብ መመገብ፣ጸሃይ በተገቢው አለማግኘት	list	list	ቫይታሚን ዲ ያነሰው ምግብ መመገብ፣ጸሃይ በተገቢው አለማግኘት	1	1
የጥፍር መንከስ ችግር የሚያስከትላቸው ችግሮች ዘርዝር	የጥፍር መጥምጥ ህመም፣የጥርስ መቦርቦር	list	list	የጥፍር መጥምጥ ህመም፣የጥርስ መቦርቦር	1	1

ጥፍር ሲሰነጠቅ እና ሲሰረጎድ ምን ያመለክታል?	ጥፍር ሲሰነጠቅ እና ሲሰረጎድ ሰውነት ውስጥ ያለ የብረት መጠን ማነስ፣ የደም ማነስ እና ከልብ ጋር የተያያዘ ችግር እንዳለ ማሳያ ይሆናል።	description	description	ጥፍር ሲሰነጠቅ እና ሲሰረጎድ ሰውነት ውስጥ ያለ የብረት መጠን ማነስ፣ የደም ማነስ እና ከልብ ጋር የተያያዘ ችግር እንዳለ ማሳያ ይሆናል።	1	0.9
የሰዉነት መቆጣት (አለርጂ) ማለት ምን ማለት ነው?	አንዳንድ ሰዎች ለተለያዩ ነገሮች ሲጋለጡ ሰዉነታቸዉ ይቆጣል። ያ ደግሞ በልዩ ልዩ መንገድ ይገለፃል። በሕክምናዉ አገላለፅ አለርጂ የሚባለዉ ማለት ነዉ።	definition	definition	አንዳንድ ሰዎች ለተለያዩ ነገሮች ሲጋለጡ ሰዉነታቸዉ ይቆጣል። ያ ደግሞ በልዩ ልዩ መንገድ ይገለፃል። በሕክምናዉ አገላለፅ አለርጂ የሚባለዉ ማለት ነዉ።	1	1
የአይን አለርጂ ምልክቱ ምንምን ነዉ?	በብዛት ከሚታዩት የአይን አለርጂ ምልክቶች ዉስጥ የአይን መቅላት፣ዉሃ የመሰለ የአይንፈሳሽና ሁለቱንም አይን ማሳከክ ናቸዉ።	list	list	በብዛት ከሚታዩት የአይን አለርጂ ምልክቶች ዉስጥ የአይን መቅላት፣ዉሃ የመሰለ የአይንፈሳሽና ሁለቱንም አይን ማሳከክ ናቸዉ።	1	0.928571

E. Conclusions

In this study, we introduced Amharic question type classification by annotating 2690 question category pairs. This overcome the scarcity of QA resources in low-resource languages like Amharic in the health domain. We compile the dataset from approximately 1,200 Amharic health documents. To our best knowledge, it is the first time that Amharic question type classification dataset released in health domain. We investigated Amharic question type classification using two experimental settings, Question Classification (QC) and Question–Answer Classification (QAC), to assess the impact of incorporating answer-aware contextual information. Two deep learning approaches were evaluated: AmharicWord2Vec combined with BiLSTM and attention, and BERT. The experimental results demonstrate that the inclusion of answer information significantly improves classification performance across both models, confirming that enriched semantic context helps the classifier better distinguish between closely related question categories. Among the evaluated models, the BERT-based QAC approach achieved the highest overall performance**,** highlighting the effectiveness of transformer-based contextual representations for low-resource language processing tasks. The Word2Vec-BiLSTM-Attention model also produced competitive results, demonstrating that recurrent neural architectures remain strong alternatives when combined with pre-trained embeddings and attention mechanisms. Furthermore, the QAC setting substantially reduced confusion between semantically similar classes, particularly definition and description, showing that answer context provides additional semantic cues for clearer category separation. The findings confirm that answeraware contextual enrichment is a valuable strategy for improving Amharic question type classification and can contribute to the development of more accurate Amharic question answering systems. This work provides an

important step toward advancing natural language processing research for low-resource languages. Future work may explore larger transformer architectures, domain-specific Amharic pre-trained language models.

Data Availability Statement: The data samples and detailed codes can accessed by contacting the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: Berhanu Bogale draft carried out the experiments and write the manuscript. Tesfa Tegegne review and supervise the QA system with a significant discussions and guidance. Solomon Teferra review the documents with a significant discussions and guidance. Gebeyehu Belay review the manuscript with significant discussions. Tesfa Tegegne, Solomon teferra and Gebyehu Belay reviewed, edited and provided feedback on each and every process of the study.

Funding: There is no funds for this study

F. References

[1] V. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni and T. R. täschel Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, "Retrievalaugmented generation for knowledge-intensive nlp tasks," Adv. Neural Inf. Process. Syst., vol. 3, 2020.

[2] S. Agrawal and N. Mishra, "Question Classification for Health Care Domain Using Rule Based Approach," in Innovative Data Communication Technologies and Application, J. S. Raj, A. Bashar, and S. R. J. Ramson, Eds., Cham: Springer International Publishing, 2020, pp. 410–419.

[3] I. H. Sarker, H. Janicke, M. A. Ferrag, and A. Abuadbba, "Multi-aspect rule-based AI: Methods, taxonomy, challenges and directions towards automation, intelligence and transparent cybersecurity modeling for critical infrastructures," Internet of Things, vol. 25, p. 101110, 2024, doi: https://doi.org/10.1016/j.iot.2024.101110.

[4] P. A. Bibin and P. Babu Anto, "Malayalam Question Answering System Based on a Deep Learning Hybrid Model of CNN and Bi-LSTM Approach," Int. J. Mod. Educ. Comput. Sci., vol. 16, no. 3, pp. 39–55, 2024, doi: 10.5815/ijmecs.2024.03.04.

[5] K. Komiya, Y. Abe, H. Morita, and Y. Kotani, "Question answering system using Q & A site corpus Query expansion and answer candidate evaluation," Springerplus, vol. 2, no. 1, p. 396, 2013, doi: 10.1186/2193-1801-2-396.

[6] B. Wang, Y. Lei, T. Yan, N. Li, and L. Guo, "Recurrent convolutional neural network: A new framework for remaining useful life prediction of machinery," Neurocomputing, vol. 379, pp. 117–129, 2020, doi: https://doi.org/10.1016/j.neucom.2019.10.064.

[7] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the Domain Adaptation of Retrieval Augmented

Generation (RAG) Models for Open Domain Question Answering," Trans. Assoc. Comput. Linguist., vol. 11, pp. 1–17, 2023, doi: 10.1162/tacl_a_00530.
[8] Q. Nguyen et al., "Advancing Question-Answering in Ophthalmology With RetrievalAugmented Generation: Benchmarking Open-Source and Proprietary Large Language Models.," Transl. Vis. Sci. Technol., vol. 14, no. 9, p. 18, Sep. 2025, doi: 10.1167/tvst.14.9.18.
[9] A. Oita, E. Filippova, and U. Toshpulatov, "A Deep Learning Question Answering System for Health Professionals at the Point of Care: A Systematic Review," J. Am. Med. Informatics Assoc., vol. 27, no. 9, pp. 1410–1419, 2020, doi: 10.1093/jamia/ocaa125.
[10] E. F. Dinsa, M. Das, and T. U. Abebe, "AI-based disease category prediction model using symptoms from low-resource Ethiopian language: Afaan Oromo text," Sci. Rep., vol. 14, no. 1, p. 11233, 2024, doi: 10.1038/s41598-024-62278-7.
[11] M.H.Uma K, "Data Collection Methods and Data Pre- processing Techniques for Healthcare Data Using Data Mining," Int. J. Sci. Eng. Res, vol. 8, pp. 1131–1136, 2017, doi: -.
[12] Agrawal et al., "Question Classification for Health Care Domain Using Rule Based Approach," springer, 2020.
[13] Naresh Kumar, Nripendra Narayan Das, Deepali Gupta, Kamali Gupta, and Jatin Bindra, "Efficient Automated Disease Diagnosis Using Machine Learning Models," 2021.
[14] M. Derje, "Amharic Factoid Question Answering System," 2020.
[15] D. S. Sachan, S. Reddy, W. Hamilton, C. Dyer, and D. Yogatama, "End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering," Adv. Neural Inf. Process. Syst., vol. 31, no. NeurIPS, pp. 25968–25981, 2021.
[16] J. Sawatzki, T. Schlippe, and M. Benner-wickner, "Deep Learning Techniques for Automatic Short Answer Grading : Predicting Scores for English and German Answers Deep Learning Techniques for Automatic Short Answer Grading : Predicting Scores for English and German Answers," no. April, pp. 0–11, 2022, doi: 10.1007/978-981-167527-0.
[17] Abreham, "Designing Amharic Question Answering model for Healthcare Using Deep Learning Approach," 2022.
[18] A. Roy, K. Sarkar, and C. Mandal, Bengali Text Classification: A New multi-class Dataset and Performance Evaluation of Machine Learning and Deep Learning Models. 2023. doi: 10.21203/rs.3.rs-3129157/v1.
[19] X. Li and D. Roth, "Learning Question Classifiers," in {COLING} 2002: The 19th International Conference on Computational Linguistics, 2002. [Online]. Available: https://aclanthology.org/C02-1150/
[20] A. Desalegn, "A Web-Based Amharic Question Answering System for Factoid Questions Using Machine Learning Approach ," - , vol., 2013.
[21] Y. A. Tilahun Abedissa, Ricardo Usbeck, "AmQA: Amharic Question Answering Dataset," 2022.
[22] X. Bai, "Text classification based on LSTM and attention," in 2018 Thirteenth International Conference on Digital Information Management (ICDIM), 2018, pp. 29– 32. doi: 10.1109/ICDIM.2018.8847061.

[23] Liu inzhang et al., “Research on Question Classification Methods in the Medical Field,” 2022.
[24] J. Liu, Y. Yang, S. Lv, J. Wang, and H. Chen, “Attention-based BiGRU-CNN for Chinese question classification,” J. Ambient Intell. Humaniz. Comput., 2019, doi: 10.1007/s12652-019-01344-9.
[25] S. Yilmaz and S. Toklu, “A deep learning analysis on question classification task using Word2vec representations,” Neural Comput. Appl., vol. 32, no. 7, pp. 2909–2928, 2020, doi: 10.1007/s00521-020-04725-w.
[26] Pranav Rajpurkar et al., “Know What You Don’t Know: Unanswerable Questions for SQuAD,” Comput. Lang., 2018.
[27] R. Chada and P. Natarajan, “FewshotQA: A simple framework for few-shot learning of question answering tasks using pre-trained text-to-text models,” EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc., pp. 6081–6090, 2021, doi: 10.18653/v1/2021.emnlp-main.491.
[28] Tadesse Destaw et al., “Question Answering Classification for Amharic Social Media Community Based Questions,” 2022.
[29] A. Sherstinsky, “Fundamentals of Recurrent Neural Network (RNN) and Long ShortTerm Memory (LSTM) network,” Phys. D Nonlinear Phenom., vol. 404, p. 132306, 2020, doi: https://doi.org/10.1016/j.physd.2019.132306.
[30] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP}), A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. doi: 10.3115/v1/D14-1181.
[31] N. D. Hailemariam, B. Guda, and T. Tefferi, “AmaSQuAD: A Benchmark for Amharic Extractive Question Answering,” 2025, [Online]. Available: http://arxiv.org/abs/2502.02047
[32] J. El Boukkouri, H.; Ferret, O.; Lavergne, T.; Noji, H.; Zweigenbaum, P.; Tsujii,
[33] “CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters.,” Proc. 28th Int. Conf. Comput. Linguist., 2020.
[34] S. W. Goitom Ybrah, “Deep Learning Based Amharic Chatbot for FAQs in Universities,” ArxIV.org, 2024.
[35] Y. Wu et al., “Google ’ s Neural Machine Translation System : Bridging the Gap between Human and Machine Translation,” pp. 1–23.
[36] H. Yu, C. Yu, Z. Wang, D. Zou, and H. Qin, “Enhancing Healthcare Through Large Language Models: A Study on Medical Question Answering,” 2024 IEEE 6th Int. Conf. Power, Intell. Comput. Syst. ICPICS 2024, pp. 895–900, 2024, doi: 10.1109/ICPICS62053.2024.10797141.
[37] Tan et al., “LSTM-Based Deep Learning Models For Non-Factoid Answer Selection,” p. 1, 2015.
[38] W. Albattah and R. U. Khan, “Impact of imbalanced features on large datasets.,” Front. big data, vol. 8, p. 1455442, 2025, doi: 10.3389/fdata.2025.1455442.
[39] H. & K. Chawla, Bowyer, “SMOTE: Synthetic minority oversampling technique,” 2002. [39] R. Sutherland, L., & Glover, “Data Privacy Concerns in

AI-Based Sports Tools: Addressing Ethical and Legal Challenges.," J. Sport. Law Ethics, vol. 29(1), pp. 72–84.
[40] M. Nabeel Asim, S. Ahmed, and A. Dengel, "Word Embedding Methods BT - Artificial Intelligence for Molecular Biology: Advanced Methods and Applications," M. Nabeel Asim, S. Ahmed, and A. Dengel, Eds., Cham: Springer Nature Switzerland, 2025, pp. 1–60. doi: 10.1007/978-3-031-90454-7_1.
[41] M. P. Timo Möller, Julian Risch, "GermanQuAD and GermanDPR: Improving NonEnglish Question Answering and Passage Retrieval," Comput. Lang., 2021.
[42] W. Hou and Y. Nie, "Seq2seq-Attention Question Answering Model".
[43] A. Paszke et al., "PyTorch: an imperative style, high-performance deep learning library," in Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA: Curran Associates Inc., 2019.
[44] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Q. Liu and D. Schlangen, Eds., Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.
[45] J. Kingma, D. P., & Ba, "Adam: A Method for Stochastic Optimization," Proc. 3rd Int. Conf. Learn. Represent., [Online]. Available: arxiv:1412.6980.
[46] R. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," J. Mach. Learn. Res., vol. 15(1), pp. 1929–1958, 2014.