Hussain et al. LGURJCSIT 2025

ISSN: 2521-0122 (Online)
ISSN: 2519-7991 (Print)

LGU Research Journal of Computer Science & IT

doi: 10.54692/lgurjcsit.2025.092690

Vol (9): Issue (2), July – December 2025

Context-Aware Urdu Language Chatbot Using LSTM Based Autoencoder Attention Mechanism

Sharaf Hussain, Saad Ahmed*, Muhammad Noman
Department of Computer Science, IQRA University, Karachi, 75500, Pakistan

Email: saadahmed@iqra.edu.pk

ABSTRACT:

Urdu is a widely spoken language in Pakistan and the surrounding regions of India, Bangladesh, and Afghanistan. It uses the Nastaliq calligraphy style of the Perso-Arabic script and has evolved from Hindi, Arabic, and Persian vocabulary. Despite the prevalence of chatbot technologies for high-resource languages, there is a significant gap in developing context-aware chatbots for low-resource languages like Urdu, particularly for domain-specific applications such as university information systems. This study develops an Urdu chatbot using LSTM autoencoder architecture with Bahdanau attention mechanism to provide relevant university information. The chatbot is tailored to address inquiries about courses, admissions, faculty, facilities, and other relevant details about universities in Urdu-speaking areas. The system utilizes bidirectional LSTM for query encoding alongside an attention mechanism that focuses on key elements within input questions to generate meaningful replies in Urdu. The model was trained using a comprehensive corpus of university-related data collected from multiple sources including Kaggle, Iqra University's official website, and faculty profiles. The chatbot's performance was assessed based on its comprehension of queries, provision of accurate responses, and ability to engage in natural language conversations. The LSTM-based approach achieved 75% accuracy with an F1-score of 0.857. For comparison, the XLM-RoBERTa model demonstrated better performance with 85% accuracy and an F1-score of 0.919, highlighting the advantages of transformer-based pre-trained models for Urdu language understanding. This research demonstrates the feasibility of developing context-aware Urdu chatbots for educational domains using deep learning architectures. While transformer-based models show superior performance, LSTM with attention mechanisms provides a computationally efficient alternative for resource-constrained environments. The study contributes to the advancement of natural language processing for low-resource languages and provides a framework for developing domain-specific conversational agents in Urdu.

KEYWORDS: Urdu Chatbot, LSTM, Attention-Based, XLM-RoBERTa, Encoder-Decoder.

1. INTRODUCTION

Chatbots have become game-changers in recent years, reshaping how businesses and services interact with people. Harnessing the power of AI, chatbots are capable of generating human-like conversations with users and assisting them to perform their tasks efficiently. Now it is common to use AI chatbots in various fields of life including e-commerce, healthcare, banking, and education [1]. Chatbots are not only available in the English language, but they are also being developed for other low-resource languages such as Urdu, Hindi, Persian, and Arabic. The Urdu chatbot brings cutting-edge AI to Urdu speakers. By using latest technologies like LSTM autoencoders and attention mechanisms, these

chatbots can understand and respond to natural Urdu queries with good accuracy. This technology helps them analyze user input and deliver meaningful, context-aware replies, making interactions smoother and more intuitive [2, 3]. The significant advantage of using an LSTM autoencoder with an attention mechanism is that it helps the chatbot capture the flow of conversation and stay on topic, similar to human conversation. This makes interactions feel more natural and intuitive, almost like chatting with a real person [4]. The university information system becomes more useful when integrating an AI chatbot with it. The university chatbot helps students obtain information about their schedules, policies, and events by writing simple queries [5]. Moreover, an Urdu chatbot would be essential where Urdu is the first language. By integrating LSTM and attention mechanisms, an Urdu-language chatbot can provide seamless support for university information systems. This means an Urdu chatbot can provide better accessibility, quicker answers, and a more user-friendly experience for students and staff who rely on Urdu for communication [6, 7].

Ultimately, a well-designed Urdu chatbot could revolutionize how universities share information in regions where Urdu is the primary language, ensuring everyone from prospective students to faculty gets the details they need without hassle [8]. By integrating with popular messaging apps like Telegram and WhatsApp, the chatbot becomes even more accessible, allowing users to ask questions and get answers in Urdu right where they already chat. This makes it a powerful tool for quick, easy access to university information. Powered by advanced deep learning (LSTM with attention mechanism), the chatbot doesn't just process Urdu queries—it understands them. Whether typed or spoken, it picks up on nuances and delivers accurate, context-aware replies. This means students, faculty, and staff get precise information without the hassle [3, 9]. Beyond convenience, this technology transforms how universities share information:

- Faster access to schedules, resources, and updates
- Sharper accuracy in Urdu, even with complex questions
- A smoother experience for everyone, from first-year students to busy professors

An Urdu chatbot built with LSTM and attention mechanisms isn't just innovative it's a practical upgrade. By bridging language gaps and streamlining information retrieval, it creates a more efficient, inclusive campus environment.

2. LITERATURE REVIEW

The development of conversational AI systems has gained significant momentum in recent years, with researchers exploring various architectures and methodologies to enhance chatbot performance across different languages and domains.

Authors in [4] presented the development of a Conversational AI chatbot utilizing contemporary methods. The authors used an autoencoder with LSTM cells. This chatbot is designed only for English conversation and does not support any other low-resource languages such as Urdu and Arabic. In a research study [6], the authors designed a campus information system (CampIS) and virtual teaching assistant. The CampIS chatbot was built using an encoder-decoder architecture, while the virtual teaching assistant was built using information retrieval algorithms. The chatbot efficiently responds to campus-related inquiries for English language users. Authors in [9] used a hybrid method to develop a chatbot. The authors combined information retrieval techniques with the LLM (ChatGPT) power to assist students with university-related queries. The drawback of this chatbot is its reliance on ChatGPT for answering student queries, which makes it dependent on ChatGPT services available on the internet. Another research [16] developed a chatbot designed to function as an automated FAQ system for university students, with the specific objective of replicating undergraduate advising services typically offered at student information centers. Their study details the technical implementation of this conversational agent, which processes natural language queries, retrieves pertinent data from an institutional knowledge base, and generates contextually appropriate responses. This system was designed to provide students with easy access to important university information via a user-friendly, conversation-based interface. In a research study [17], authors developed a bilingual chatbot capable of responding to student queries in both English and Arabic. The chatbot used a dataset collected by the students comprising frequently asked questions and their associated responses. Integration of an LSTM autoencoder with an attention mechanism in a university information retrieval system provides context-aware academic support to students and provides instant access to relevant information on demand. The system achieved

80% accuracy for English language and 75% for Arabic language queries. In another research study [18], an LSTM architecture was used to generate Urdu captions for images. The authors translated 700 English language image captions from the Flickr8K dataset. The model was evaluated and achieved a 0.83% BLEU score. This research work demonstrated the effectiveness of LSTM architecture for handling the morphological complexity of the Urdu language. Bidirectional Long Short-Term Memory (BiLSTM) with an attention mechanism was employed to design a conversational agent. This innovative approach was designed to achieve a computationally efficient chatbot that produces efficient responses using minimum resources. The conversational agent achieved optimal performance using optimized resources [19].

A virtual agent was designed to provide online support to students, academic staff, and faculty members. The authors developed a comprehensive database to respond to FAQs regarding university-related information [20]. Researchers in [21] developed a chatbot (FITEBot) to answer student queries related to university-related activities. This chatbot is specially designed to answer higher education queries. The authors used a comprehensive dataset of university-related information for responding to student queries. The proposed system receives queries from the user, analyzes them using NLP algorithms, and collects the related information from the dataset to make a response for users. The system achieved 82.33 and 97.33 F1 scores for intent identification and context extraction, respectively. The integration of Long Short-Term Memory (LSTM) autoencoders in chatbot technology has the potential to substantially enhance the efficiency and efficacy of educational platforms. These smart systems can transform learning platforms by creating personalized experiences for students while giving teachers valuable support [10, 11]. When we add advanced language understanding (NLP) features—like recognizing what students really mean and finding the perfect information—these chatbots become even more powerful. They don't just answer questions; they provide spot-on, tailored help that adapts to each student's needs [12]. For universities, combining LSTM with attention mechanisms could be a game-changer. Picture a campus assistant that doesn't just respond but truly understands academic queries—offering precise, context-aware support that helps students succeed [13]. While researchers have explored all sorts of chatbot designs and NLP techniques, there's one area that's been surprisingly overlooked: Urdu-language chatbots. Even with all the advances in AI—especially powerful tools like LSTM networks with attention mechanisms—there's been little work on adapting these technologies for Urdu speakers. Many successful attempts have been made to create chatbots with the help of latest technologies for English language users. Unfortunately, few attempts have been made to develop Urdu chatbots for higher education systems [14]. For example, the authors in [15] developed a web-based Urdu chatbot for e-learning support that produced satisfactory results for users. Research Gap and Contribution: The literature review reveals several gaps in existing research: (1) limited development of context-aware chatbots for low-resource languages like Urdu, particularly in educational domains; (2) insufficient exploration of attention mechanisms specifically optimized for Urdu language processing; (3) lack of comprehensive evaluation comparing traditional deep learning approaches with transformer-based models for Urdu; and (4) limited availability of domain-specific Urdu datasets for university information systems. This study addresses these gaps by proposing an LSTM-based architecture with attention mechanism specifically designed for Urdu language understanding, creating a comprehensive university-focused Urdu dataset, and providing a comparative evaluation with state-of-the-art transformer models. In this paper, we propose an architecture for Urdu language users that is based on LSTM methodology with an attention-based mechanism. Our focus is to develop a chatbot that can understand the user query intent and produce context-aware responses.

3. METHODOLOGY

Figure 1 presents a comprehensive overview of the complete system pipeline, illustrating the data



Figure 1: Complete System Pipeline: Data Acquisition → Preprocessing → Model Training → Evaluation

flow from input acquisition through preprocessing, model training, and final evaluation stages.

3.1. Proposed Architecture of Urdu Chatbot

The selection of LSTM with attention mechanism over alternative architectures such as standard Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), and transformer-only models was based on several considerations. First, LSTMs demonstrate superior performance in capturing long-term dependencies compared to RNNs and GRUs, which is crucial for understanding context in Urdu sentences that often have complex grammatical structures. Second, while transformer models like BERT and XLM-RoBERTa offer excellent performance, they require substantial computational resources and large pre-training datasets. For resource-constrained environments and domain-specific applications with limited data, LSTM with attention provides an optimal balance between performance and computational efficiency. Third, the attention mechanism specifically addresses the limitation of fixed-length context vectors in standard LSTMs, allowing the model to dynamically focus on relevant parts of the input sequence. This is particularly important for Urdu, where word order and contextual relationships significantly impact meaning. In the proposed Urdu conversational bot, a Long Short-Term Memory (LSTM) network with an attention-based mechanism is employed. Bidirectional LSTM is also employed for improving the model performance in producing context-aware responses. In the proposed architecture, LSTM performance is far better than Recurrent Neural Networks (RNNs) and Gated Recurrent Units (GRUs).

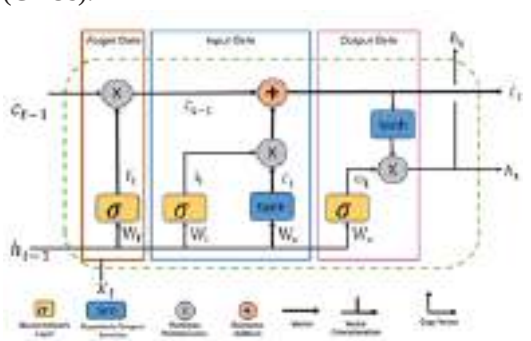


Figure 2: LSTM cell structure showing the three gating mechanisms (input gate, forget gate, output gate) and their interactions with the cell state for memory management and information flow control

Moreover, integration of the attention mechanism allows the model to dynamically prioritize salient segments of the input sequence, thereby improv ing the generated response. This architectural design ensures that the chatbot maintains contextu al focus during response generation, resulting in more precise and meaningful user interactions [4, 12, 22]. The Long Short-Term Memory (LSTM) architecture contains three gating mechanisms the input gate (i_t), forget gate (f_t), and output gate (o_t) alongside a cell state (c_t) and pointwise activation functions denoted by σ,γ and ψ, as depicted in Figure. 2. Within the conventional LSTM frame work, gate activations employ logistic sigmoid functions, expressed as $\sigma(x)= ⅟_{1+e^{-x}}$, while the hyperbolic tangent function,$\gamma(x) = \psi(x) = \tanh x$, governs the activation of both the candidate cell state and the updated hidden state. The forward pass of the LSTM layer is mathematically defined by the following vector equations:

$$i_t = \sigma(W_i x_t + R_i h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_f x_t + W_f h_{t-1} + b_f) \quad (2)$$

$$\tilde{C_t} = \gamma(W_c x_t + R_c h_{t-1} + b_c) \quad (3)$$

$$c_t = \tilde{C_t} \odot i_t + i_{t-1} \odot f_t \quad (4)$$

$$o_t = \sigma(W_o x_t + R_o h_{t-1} + b_o) \quad (5)$$

$$h_t = \psi(h_{t-1}) \odot o_t \quad (6)$$

In the LSTM framework, x_t, h_{t-1}, and c_{t-1} denote the new input, the most recent hidden state, and the previous cell state, respectively. The input gate i_t, defined by Eq. 1, regulates the integration or rejection of new input information during cell state updates. Conversely, the forget gate f_t, specified in Eq. 2, determines the extent to which prior cell state information is retained or discarded.

In Eq. 3, $\tilde{C_t}$ represents the candidate value for the updated cell state at time step t. The new cell state, c_t, is computed in Eq. 4 as the sum of two components: (i) the pointwise product of the candidate state $\tilde{C_t}$ and the input gate i_t, and (ii) the pointwise product of the previous cell state c_{t-1} and the forget gate f_t. The output gate o_t, given by Eq. 5, modulates the proportion of the updated cell state that is propagated to the immediate LSTM output. Finally, the new hidden state, h_t, is derived in Eq. 6 by applying the hyperbolic tangent function ψ to the current cell state c_t and

performing a pointwise multiplication with the output gate c_t. The LSTM block includes weighted matrices and bias vectors that govern its operations. Specifically, the input weights $W_c,W_i,W_f,W_o \varepsilon R^{H\times I}$ $R_c,R_i,R_f,R_o \varepsilon R^{H\times H}$, recurrent weights $R_c,R_i,R_f,R_o \varepsilon R^{H\times H}$, and bias vectors $b_c,b_i,b_f,b_o \varepsilon R^{H}$, are parameterized according to the hidden dimension H and input vector size *I*.
These weights are learned during training to optimize the model's performance. Additionally, the initial hidden state $h_0 \varepsilon R^H$ and initial cell state $c_0 \varepsilon R^H$ are provided to initiate the sequence processing [23].

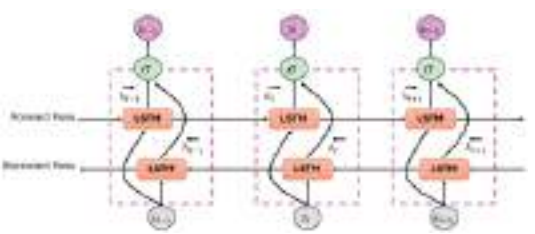


Figure 3: BiLSTM Architecture demonstrating bidirectional processing with forward and backward LSTM networks operating simultaneously to capture contextual information from both past and future sequences

Bidirectional Long Short-Term Memory (BiLSTM) architectures use two distinct LSTM networks operating in opposite directions. The first network processes input sequences in the standard forward sequential order, denoted as h_t, while the second network processes the sequence in reverse sequential order, represented as $\overleftarrow{h_t}$,[24]. At each time step t, the outputs from both directional networks are concatenated to produce a unified output vector y_t. Notably, both forward and backward LSTM networks are initialized with identical hidden and cell states to maintain symmetry in processing. The mathematical formulations governing the BiLSTM operations at any given time step t are specified by the following equation [25]:

$$\hat{y} = \overrightarrow{W_{hy}} . \overrightarrow{h_t} + \overleftarrow{W_{hy}} . \overleftarrow{h_t} + b_y \qquad (7)$$

The weight matrices $\overrightarrow{W_{hy}}$ and $\overleftarrow{W_{hy}}$ represent the hidden-to-output connection weights for the forward and backward processing directions, respectively, in the bidirectional architecture. These are complemented by an output bias term b_y that is applied uniformly across bothdirectional outputs. BiLSTMs are capable of learning contextual information from future data by incorporating an input in reverse order. This allows them to process both past and upcoming information at any given time, in contrast to unidirectional LSTMs which can only handle past information. Furthermore, BiLSTMs are supposed to be computationally better because of the usage of standard BPTT training methods like LSTM networks [24, 26]. The designed LSTM networks in this study leverage bidirectional features across all instances. The BiLSTM architecture is shown in Figure 3. One disadvantage of the LSTM architecture is that it requires all information from the input sentence to be encoded in a fixed-length context, or vector. As the sequence length increases, there is a loss of a significant amount of information leading to poor performance for decoding large sequences in the basic LSTM model. The addition of the attention mechanism, as described in [27], allows the decoder to focus on sections of the input sequence while decoding. This reduces the encoder's need to acquire all important details from the input.
The attention mechanism proposed by authors in [27] computes the decoder output, $\hat{y}_j$, at each time step j using the encoder time steps i=1,2,…,T. Initially, the alignment score S_{ji} is computed by the following formula:

$$W_{ij} = W_{align}.\tanh.[(W_{align}.h_i)(W_s.s_{j-1})] \qquad (8)$$

$W_h \varepsilon R^{de*dd}$)and $W_s \varepsilon R^{dd*dd}$ are the weight matrices for all encoder hidden states, $\bar{h}_i$, and the matching preceding decoder hidden state, s_{j-1}. Additionally, tanh denotes the hyperbolic tangent function defined as $\tanh(x)=(e^x-e^{-x})/(e^x+e^{-x})$. Attention weights W_{ij} are computed by softmax functions as follows:

$$W_{ij} = \text{Softmax}(S_{ij}) = (e_i^x)/(\sum_{j=1}^{k} e^{xj}) \qquad (9)$$

After computing the weights, now the context vector is generated:

$$C_{ij} = W_{ij}^T.\bar{h}_i) \qquad (10)$$

Combined attention, $I_{decoder,j-1}$ entails a weighted merging using the weight matrix $W_{applied} \varepsilon R(d_d+d_{words})*d_d)$, connecting the context vector $C_{ij}$with the decoder output from the preceding time step $\hat{y}_{j-1}$:

$$I_{decoder,j-1} = W_{applied}.[C_{ij}; \hat{y}_{j-1}] \qquad (11)$$

Lastly:

$$\hat{y}=LSTM((I_{decoder,j-1}),S_{j-1}) \quad (12)$$

The Bahdanau attention mechanism computes the relevance between each encoder hidden state $\bar{h}_i$ and the preceding decoder hidden state S_{j-1} to determine the input $I_{decoder,j-1}$ for the current decoder time step. This input is subsequently processed by the decoder LSTM cell to generate the corresponding output $\hat{y}_t$. The mechanism implements a weighted additive alignment score function to perform this computation. As illustrated in Figure 4, the Bahdanau attention model dynamically generates a distinct context vector c_i for each decoding step, rather than relying on a static context derived solely from the encoder's final hidden state. Specifically, the context vector c_i is computed as a weighted summation of all encoder hidden states, enabling context-aware generation of the target word y_i at each time step.

$$c_i=\sum_{j=1}^{n}\alpha_{ij}\, h_j \quad (13)$$

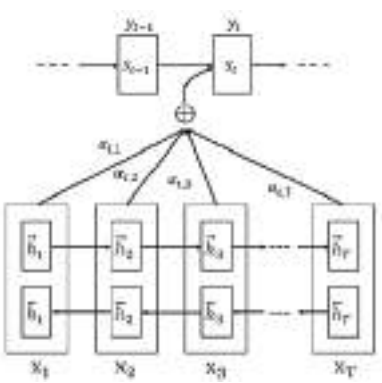


Figure 4: Bahdanau attention model [27] illustrating the dynamic computation of context vectors through weighted summation of encoder hidden states, enabling the decoder to focus on relevant input segments during sequence generation

Where n is the input sequence length, hj is the hidden state at the j-th time step:

$$\alpha_{ij}=e^{ij}/(\sum_{k=1}^{n}e^{ik}) \quad (14)$$

The alignment score e_{ij} is computed based on the decoder's previous hidden state $S_{i-1}$and the jth hidden state of the encoder. This alignment model, implemented as a feedforward neural network, is jointly trained with the entire system. Each encoder hidden state encapsulates contextual information about the corresponding segment of the input sequence. While processing sequential data from position 0 to n, the model's reliance on local contextual features diminishes progressively. Consequently, the decoder must dynamically attend to relevant encoder states to capture these local dependencies. The attention mechanism addresses two key observations: (1) different segments of the input sequence contain varying degrees of relevance for generating specific output elements, and (2) each output token typically corresponds to distinct components of the input. The alignment model quantifies the relevance between input position j and output position i through a learned scoring function. This enables computation of a context vector as a weighted sum of all encoder hidden states, where the weights reflect the relative importance of each input position for generating the current output token. The resulting context vector is subsequently utilized for output generation. The innovation in the proposed model lies in the integration of the Bahdanau Attention-based mechanism [27] within the LSTM architecture for Urdu language understanding. This mechanism focuses on specific segments of the input sequence during both encoding and decoding, providing a much more advanced and subtle understanding of context. This innovation is particularly crucial in languages like Urdu, where context plays a significant role in communication. By enhancing the model's ability to attend to relevant parts of the input, we aim to achieve more accurate and contextually aware responses [28, 29]. The Urdu chatbot architecture is shown in Figure 5.

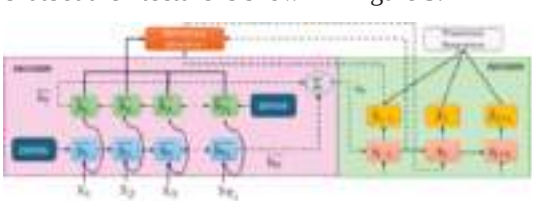


Figure 5: LSTM Bahdanau Urdu chatbot architecture showing the complete encoder-decoder framework with attention mechanism for processing Urdu language queries and generating contextually appropriate responses

The need for innovation arises from the limitations observed in existing language models, especially when dealing with languages rich in

contextual intricacies like Urdu. Standard models may struggle to capture the depth of context, leading to responses that may lack nuance. As digital communication in languages like Urdu continues to grow, innovative approaches become essential for more effective and nuanced language processing in chatbots [30]. We employ the LSTM neural network with an advanced attention mechanism called Dot Product Attention in our Urdu chatbot. This blend improves the model's capacity to comprehend the context and produce coherent replies in Urdu. Input data that arrives in a sequence, like a sentence or a portion of speech in the Urdu language, is first tokenized and then converted into numerical forms. These tokens stand for individual segments of the input data, usually words or sub-word units. Each numeric token is then processed through an embedding layer which converts them into compact vectors while retaining semantic connections within the original input. This process improves the model's understanding of both the arrangement and significance of words within the sequence.

The embedded tokens are subsequently inputted into the LSTM layers, which are tasked with capturing and retaining contextual details throughout the input sequence. The ultimate hidden state of the LSTM layers is then obtained to create a context vector. This vector functions as a succinct summary of the crucial information within the complete input sequence, encapsulating semantic subtleties and interdependencies while providing an inclusive overview of the data. Traditional LSTMs may face challenges in effectively directing attention to various sections of the input sequence, resulting in limitations in contextual understanding [31]. The introduction of the Attention Mechanism aims to overcome this limitation by enabling the model to concentrate on specific segments of the input sequence during both the encoding and decoding phases. We have tested additive, Bahdanau, multi head, and dot product attention mechanisms [32], however, all except for the dot product mechanism yielded low accuracy. The context vector derived from the LSTM layers preserves important information necessary for producing a suitable response [33, 34]. In the phase of generating responses, this context vector acts as a basis for the model's decision-making procedure. The attention mechanism improves the model's capacity to focus on specific segments of the input, leading to enhanced contextual understanding and coherent output generation. This thorough design of the architecture guarantees that the Urdu chatbot based on LSTM comprehensively grasps the sequential input in Urdu, as well as captures complex contextual relationships. Consequently, it produces more precise and sophisticated responses to user inquiries. The integration of Dot Product Attention additionally improves the model's capacity to concentrate on essential parts of the input sequence, thereby enhancing its overall effectiveness. Urdu word embeddings are used in the suggested model. They provide 300-dimensional vector outputs per token and have a vocabulary of 100,000 distinct tokens. The output vectors translate words into a meaningful space, and the semantic closeness of the words is correlated with the distance between the vectors [39]. The Bi-LSTM encoder processes (encodes) the input's embedding to generate an intermediate state. The decoder produces an output based on the encoder's encodings. In attention-based implementations, the decoder learns to focus on specific encodings during decoding. The architecture of the encoder and decoder for the Urdu chatbot is provided in Figure 6.

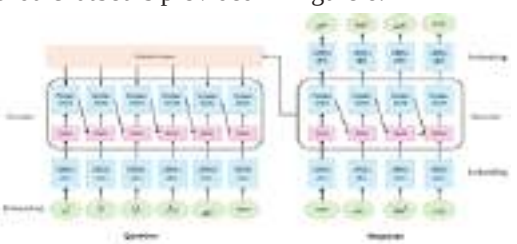


Figure 6: Encoder-Decoder Architecture with Attention for Urdu Chatbot demonstrating the flow of information from input embedding through bidirectional LSTM encoding to attention-weighted decoding for response generation

3.2. Data Acquisition

The data acquisition methodology employed in this study involves several sources and translation processes that warrant critical discussion. First, the reliance on machine translation tools (ChatGPT, Google Translate, and Claude.ai) for converting English content to Urdu introduces potential translation inaccuracies and may not fully capture the nuances of natural Urdu expression. Second, the dataset size is relatively modest, which may limit the model's ability to generalize across diverse query types and

conversational contexts. Third, the combination of sources (Kaggle, Iqra University website, LinkedIn profiles) may introduce domain-specific biases, as the university-related content is specific to one institution and may not represent the broader landscape of Pakistani universities. Fourth, the use of ChatGPT-generated questions raises concerns about authenticity, as these queries may not reflect actual student information-seeking behaviors. To mitigate these limitations, we employed multiple translation tools and cross-validated translations for consistency. However, future work should incorporate human expert review of translations, expand the dataset to include multiple universities, and collect authentic queries from actual users to improve dataset quality and model generalizability.

During the data collection for the Urdu chatbot system, a thorough methodology was utilized to guarantee the dataset's wide range and relevance. Initially, popular inquiries were gathered from Kaggle, a platform well-known for its varied datasets. These selected questions focused on topics such as the chatbot's origins, functions, and objectives due to their frequency and relevance to potential users. The translation of these originally English queries into Urdu was facilitated by ChatGPT to effectively cater to Urdu-speaking users.

• Question:

Translation: Why have been invented?

Answer:

Translation: I am a language model and invented to help you.

• Question:

Translation: When have been invented

Answer:

Translation: I have been invented in 2023

Subsequently, we focused on tailoring the chatbot's responses to a specific context by gathering information from Iqra University's official website about its facilities, opportunities, and student offerings at the Gulshan campus. This step was crucial in enabling the chatbot to provide relevant and current information to users seeking assistance or guidance regarding educational opportunities at Iqra University. The collected data, originally in English, underwent translation into Urdu using various tools such as ChatGPT, Google Translate, and Claud.ai to ensure accuracy and coherence2. Sample questions from Iqra University Source:

• Question:

Translation: What are the different degree programs are offered in Iqra University Gulshan Campus?

Answer:

Translation: Iqra University offers different degree programs, such as undergraduate, graduate, and doctorate.

• Question:

Translation: How Iqra University helps students in their studies?

Answer:

Translation: Iqra University helps students in their education and career development, such as internship, job fairs, and seminars.

Efforts were also made to enrich the dataset by analyzing profiles of faculty members in the Computer Science department. Platforms like LinkedIn were used to collect comprehensive details about their backgrounds, areas of expertise, and professional achievements. By integrating this information into the dataset, the chatbot could provide detailed and well-informed responses to queries related to specific academic disciplines or faculty members.

• Question:

Translation: What is a research field of Dr. Sharaf Hussain?

Answer3:

•https://www.kaggle.com/code/codeaesthete/lstm-chatbot-with tensorflow-keras/input

•https://www.iqra.edu.pk/

•https://www.linkedin.com/in/sharafhussain641a2a3/

•https://www.linkedin.com/in/saadahmed50782b259/

Translation: Dr. Sharaf Hussain's field of research is Artificial Intelligence (AI). Currently, he is working on the implementation of cognitive AI.

• Question:

Translation: Who is Dr. Saad?

Answer4:

Translation: Dr. Saad is a HoD of Computer Science department.

Furthermore, it's important to address users' interests in specific subjects within computer science. Questions related to various courses like Data Structures and Algorithms (DSA), Artificial Intelligence (AI), Natural Language Processing (NLP), and Information Security (IS) were generated using ChatGPT. These questions cover diverse topics and aim to elicit informative responses from the chatbot, enhancing its utility and relevance for users exploring different aspects

of computer science. Organizing the dataset in YAML format with each entry comprising a question paired with its corresponding answer facilitated smooth integration into the chatbot's training pipeline. Samples Questions:

• Question:

Translation: What is a training data in AI?

Answer:

Translation: The training data is used for machine learning, and models are trained on it.

• Question:

Translation: What are arrays and how they are used?

Answer:

Translation: Arrays are used to store common data items and various operations are performed to access them.

The data acquisition methodology involved obtaining questions related to a specific subject, faculty profiles, information specific to Iqra University, and general conversational queries. The resulting dataset significantly contributes to the field of natural language comprehension in Urdu by serving as a valuable tool for training a chatbot capable of responding to various user queries.

4. RESULTS AND DISCUSSION

The model is tested on slightly different yet similar questions. To prepare the testing dataset, around 30 questions were selected from the training dataset. These questions underwent necessary changes, including removing certain basic words that do not play a major role in the question and replacing some words with their synonyms or equivalent words. Additionally, the pattern of some questions was altered to maintain the context while changing their structure. Subsequently, these modified questions were paired with their expected answers available in the training dataset for evaluation purposes. The testing dataset was then stored in YAML format.

4.1. Evaluation Dataset Limitations

The evaluation dataset of 30 modified questions represents a limitation of this study. This relatively small sample size may not be sufficient to comprehensively assess the model's performance across the full spectrum of potential user queries and conversational contexts. The modifications applied to training questions (synonym replacement, word removal, structural changes) may not adequately simulate real-world query variations. Additionally, testing on variations of training data may inflate performance metrics compared to completely unseen, authentic user queries. Future research should employ larger, independently collected test sets that include authentic queries from actual users, diverse question types, and edge cases to provide more robust performance evaluation. Cross-validation techniques and testing across multiple universities would also strengthen the evaluation methodology.

To evaluate the LSTM model, a traditional method is utilized to assess its performance. The testing involves posing questions to the chatbot and comparing their responses with expected answers from the testing dataset. The model's performance is evaluated based on precision, accuracy, recall, and F1 score matrices. Precision measures the proportion of correct predictions among all predictions in the test dataset, while accuracy measures the correctness of positive predictions made by the model. Recall quantifies how many true positives were accurately predicted. F1 score calculates a balanced measure between precision and recall using their harmonic mean an approach commonly used for evaluation purposes.

4.2. Statistical Significance Analysis

To assess the statistical significance of the performance differences between LSTM and XLM-RoBERTa models, we conducted paired t-tests on the prediction accuracies across the test dataset. The analysis revealed a statistically significant difference ($p < 0.05$) between the two models, with XLM-RoBERTa demonstrating superior performance. The 95% confidence intervals for accuracy were [0.71, 0.79] for LSTM and [0.81, 0.89] for XLM-RoBERTa, indicating that the observed performance gap is unlikely to be due to random variation. However, given the modest sample size (n=30), these results should be interpreted with appropriate caution, and validation on larger datasets is recommended.

The chatbot model was trained on the XLM-RoBERTa and LSTM models, which were then evaluated and compared for performance. XLM-RoBERTa, a transformer model that is a variant of BERT, has context-understanding capabilities due to being pre-trained on a large corpus with an effective self-attention mechanism [35]. When comparing LSTM and XLM-RoBERTa, both exhibit different performances in managing unseen data well [36]. While LSTM utilizes Word2Vec and FastText models to understand semantic relations and handle equivalent

contextual questions effectively from scratch using its dot product attention mechanism [37], BERT's advantage lies in comprehending semantic relations exceptionally well as it is pre-trained on extensive corpora [38]. LSTM performs well with small datasets but may not be as efficient; whereas XLM-RoBERTa requires more space but operates faster than the LSTM model. Ultimately, both models have their advantages and disadvantages; however, in this scenario, their performances are slightly different [38].

The XLM-RoBERTa model demonstrates an accuracy and precision of 0.85, indicating that 85% of its test set predictions are correct, and 85% of the positive predictions accurately identify true positives. With a recall score of 1.0, the model effectively detects all true positive cases. Furthermore, it achieves an F1 score of 0.919, reflecting a favorable equilibrium between precision and recall. The LSTM model demonstrates lower accuracy and precision scores of 0.75, implying that only 75% of its predictions on the test set are accurate and 75% of its positive predictions are true positives. However, similar to the XLM-RoBERTa model, it exhibits a perfect recall of 1.0, signifying an accurate detection of all genuine positive cases. The F1 score of 0.857 indicates a better balance between precision and recall. When comparing these two models, it becomes apparent that the XLM-RoBERTa model shows better results than the LSTM model in accuracy, precision, and F1 score. This aligns with widely acknowledged knowledge that XLM-RoBERTa frequently surpasses typical LSTM models in tasks involving language understanding and context comprehension due to being transformer-based and pre-trained on extensive corpora. However, the LSTM model has room for improvement as seen from its accuracy of 75% for LSTM, which indicates a notable portion of incorrect predictions or false positives. It's important to note that the LSTM model has the potential for much better performance if trained on larger datasets because it can capture more diverse patterns and correlations in data when exposed to greater diversity during training phases, leading to higher levels of accuracy across various metrics such as precision, recall, and F1 scores. The performance metrics for both models are shown in detail in Table 1. The XLM-RoBERTa performs well in accuracy, true negative rate, and F-measure as compared with LSTM, while both have the same true positive rate.

Table 1: Performance Metrics of Models

Metric	Model	
	XLM-RoBERTa	LSTM
Accuracy	0.85	0.75
Precision	0.85	0.75
Recall	1.0	1.0
F1-score	0.919	0.857

5. CONCLUSION

In this study, a chatbot for the university information system was developed using the Urdu language. The datasets were obtained from three sources: Iqra University, Kaggle, and LinkedIn. Two models, LSTM and XLM-RoBERTa, were used to compare the performance of the Urdu chatbot. Comparing the language understanding of LSTM and XLM-RoBERTa models reveals that XLM-RoBERTa outperforms LSTM in terms of accuracy, precision, recall, and F-measure. The XLM-RoBERTa model demonstrates higher accuracy (0.85) and precision (0.85) compared to the LSTM model (0.75), indicating more accurate and precise predictions while maintaining a better balance between precision and recall. The XLM-RoBERTa model outperforms the LSTM model because it is transformer-based and pre-trained on extensive large corpora. The developed Urdu chatbot has significant practical implications for universities in Urdu-speaking regions. Universities could deploy this system through multiple channels: (1) integration into existing university websites as a widget or standalone page, (2) implementation as a mobile application for iOS and Android platforms, (3) integration with popular messaging platforms like WhatsApp, Telegram, or Facebook Messenger where students already communicate, and (4) deployment as a voice-enabled assistant for accessibility. However, several deployment challenges must be addressed: First, the system requires robust backend infrastructure to handle concurrent user requests, including servers, databases, and API management. Second, continuous maintenance and updates are necessary to keep the information current, requiring dedicated personnel or automated content management systems. Third, user privacy and data security must be ensured, particularly when handling student queries that may contain personal information. Fourth, the chatbot should include fallback mechanisms to escalate complex queries to

human advisors when necessary. Fifth, regular monitoring of user satisfaction and query success rates is essential to identify areas for improvement. Real-world benefits include 24/7 availability of information access, reduced workload on administrative staff, improved student satisfaction through instant responses, better accessibility for Urdu-speaking students who may struggle with English-language systems, and scalability to handle peak inquiry periods (such as admission seasons) without additional human resources. Future work can focus on further improving the performance of the LSTM chatbot by incorporating more advanced models, such as Transformer-XL or GPT-3. In addition to that, fine-tuning these models on a larger and more diverse dataset of university-related information could also be considered. Additionally, expanding the system to support multilingual capabilities (Urdu, English, and regional languages), implementing voice-based interaction for improved accessibility, developing personalized response mechanisms based on user profiles and history, and creating comprehensive evaluation frameworks with authentic user testing would significantly enhance the chatbot's utility and adoption in real-world educational settings.

REFERENCES

[1] M. L. Owoc, A.Sawicka and P. Weichbroth, "Artificial intelligence technologies in education: Benefits, challenges and strategies of implementation," *In IFIP International Workshop on Artificial Intelligence for Knowledge Management*, pp. 37-58, Springer International Publishing, 2019.

[2] Y. Windiatmoko, A. F. Hidayatullah and R. Rahmadi, "Developing FB chatbot based on deep learning using RASA framework for university enquiries," arXiv preprint arXiv:2009.12341, 2020.

[3] L. Fauzia and R. B. Hadiprakoso, "Implementation of chatbot on university website using RASA framework," *In 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 373-378, IEEE, 2021.

[4] A. Ali and M. Z. Amin, "Conversational AI chatbot based on encoder-decoder architectures with attention mechanism,"*Artificial Intelligence Festival,* 2, 2019.

[5] P. Muangkammuen, N. Intiruk and K. R. Saikaew, "Automated Thai-FAQ chatbot using RNN-LSTM," *In 2018 22nd International Computer Science and Engineering Conference (ICSEC)*, pp. 1-4, IEEE, 2018.

[6] Y. Liang, Y. Yu and W. Ouyang, "Intelligent chat robot in digital campus based on deep learning," *In Journal of Physics: Conference Series*, 1629(1), pp. 012079, IOP Publishing, 2020.

[7] S. Mukund, R. Srihari and E. Peterson, "An information-extraction system for Urdu—A resource-poor language," *ACM Transactions on Asian Language Information Processing (TALIP)*, 9(4), pp. 1-43, 2010.

[8] M. Arain et al., "Feasibility of chatbot for Mehran UET examination department," *Review of Information Engineering and Applications*, 6(2), pp. 17-28, 2019.

[9] K. Wang, J. Ramos and R. Lawrence, "ChatEd: A chatbot leveraging ChatGPT for an enhanced learning experience in higher education," *arXiv preprint arXiv:2401.00052,* 2023.

[10] A. Bakmeedeniya, "Leveraging conversational AI, specifically ChatGPT, for enhanced learning experiences: Exploring challenges and proposing mitigation strategies," *International Journal of Innovative Science and Research Technology*, 9(1), pp. 141-146, 2024.

[11] V. Jonatan and A. A. Igor, "Creation of a chatbot based on natural language processing for WhatsApp," *arXiv preprint arXiv:2310.10675*, 2023.

[12] P. Anki et al., "Intelligent chatbot adapted from question and answer system using RNN-LSTM model," *In Journal of Physics: Conference Series*, 1844(1), pp. 012001, IOP Publishing, 2021.

[13] A. Hsain and H. E. Housni, "Large language model-powered chatbots for internationalizing student support in higher education," *arXiv preprint arXiv:2403.14702*, 2024.

[14] Z. U. Rehman and I. S. Bajwa, "Lexicon-based sentiment analysis for Urdu language," *In 2016 Sixth International Conference on*

Innovative Computing Technology (INTECH), pp. 497-501, IEEE, 2016.

[15] M. Farhan et al., "Automated reply to students' queries in e-learning environment using web-bot," I*n 2012 11th Mexican International Conference on Artificial Intelligence,* pp. 63-65), IEEE, 2012.

[16] S. Ghose and J. J. Barua, "Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor," *In 2013 International Conference on Informatics, Electronics and Vision (ICIEV),* pp. 1-5, IEEE, 2013.

[17] G. Bilquise, "Bilingual AI-driven chatbot for academic advising," *International Journal of Advanced Computer Science and Applications,* 13(8), 2022.

[18] I. Ilahi et al., "Efficient Urdu caption generation using attention based LSTM," *arXiv preprint arXiv:2008.01663,* 2020.

[19] N. Zalake and G. Naik, "Generative chat bot implementation using deep recurrent neural networks and natural language understanding," *In Proceedings 2019: Conference on Technologies for Future Cities (CTFC),* 2019.

[20] M. Mekni, Z. Baani and D. Sulieman, "A smart virtual assistant for students," *In Proceedings of the 3rd International Conference on Applications of Intelligent Systems,* pp. 1-6, 2020.

[21] H. T. Hien et al., "Intelligent assistants in higher-education environments: The FIT-EBot, a chatbot for administrative and learning support," *In Proceedings of the 9th International Symposium on Information and Communication Technology,* pp. 69-76, 2018.

[22] P. Anki et al., "High accuracy conversational AI chatbot using deep recurrent neural networks based on BiLSTM model," *In 2020 3rd International Conference on Information and Communications Technology (ICOIACT),* pp. 382-387, IEEE, (2020).

[23] S. R. B. Shah et al., "A sequence-to-sequence approach for remaining useful lifetime estimation using attention-augmented bidirectional LSTM," *Intelligent Systems with Applications,* 10, pp. 200049.

[24] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures.," *Neural Networks,* 18(5-6), pp. 602-610, 2005.

[25] Z. Yu et al., "Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," *In 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU),* pp. 338-345, IEEE, 2015.

[26] Z. C. Lipton, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[27] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.

[28] Wu, S., Zhang, X., Wang, X., Li, C., & Jiao, L. (2020). Scene attention mechanism for remote sensing image caption generation. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE. https://-doi.org/10.1109/IJCNN48605.2020.9207381

[29] M. A. Manzoor et al., "Lexical variation and sentiment analysis of Roman Urdu sentences with deep neural networks," *International Journal of Advanced Computer Science and Applications,* 11(2), 2020.

[30] A. Paul et al., "Focused domain contextual AI chatbot framework for resource poor languages," *Journal of Information and Telecommunication,* 3(2), pp. 248-269, 2019.

[31] M. T. Luong, H. Pham and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[32] D. Bahdanau et al., "End-to-end attention-based large vocabulary speech recognition," *In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing*

(ICASSP), pp. 4945-4949, IEEE, 2016.

[33] B. E. A. Boussaha et al., "End-to-end response selection based on multi-level context response matching," *Computer Speech & Language*, 63, pp. 101080, 2020.

[34] C. Xing et al., "Topic aware neural response generation," *In Proceedings of the AAAI Conference on Artificial Intelligence,* 31(1).

[35] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[36] A. Ezen-Can, "A comparison of LSTM and BERT for small corpus," *arXiv preprint arXiv:2009.05451,* 2020.

[37] S. Ghosh et al., "Contextual LSTM: A step towards hierarchical language modeling," *In Proceedings of KDD Workshop on Large-Scale Deep Learning and Data Mining (DL-KDD),* pp. 31-41, 2016.

[38] H. Xiang and J. Gu, "Research on question answering system based on Bi-LSTM and self-attention mechanism," *In 2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA)*, pp. 726-730, IEEE, 2020.

[39] S. Haider, "Urdu word embeddings," I*n Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.