Journal on
Artificial Intelligence

Tech Science Press

Doi:10.32604/jai.2025.070762

ARTICLE

Customer Service Support System: A Chatbot for University Reception

Muhammad Adeen Jamal[1], **Bilal Khan**[2,*], **Sameed Ur Rehman**[1] **and Wahab Khan**[1]

[1]Department of Computer Science, City University of Science and Technology, Peshawar, 25120, Pakistan
[2]Department of Computer Science, University of Engineering and Technology, Mardan, 23200, Pakistan

*Corresponding Author: Bilal Khan. Email: bilalsoft63@gmail.com

Received: 23 July 2025; Accepted: 19 September 2025; Published: 20 October 2025

ABSTRACT: The development of artificial intelligence (AI) has sparked the invention of chatbots, which are intelligent conversational agents. These chatbots have the potential to completely transform how people interact while enhancing user experience. This study explores the building along with its execution of a chatbot for customer service support at a university reception using recurrent neural networks (RNNs). To increase user requests, the accuracy of the information, and overall satisfaction with the service, it evaluates machine learning models including RNN, XLNet, and Bidirectional Encoder Representations from Transformers (BERT). In this research project, data were gathered from university offices and students, documenting an array of daily questions that frequently arise at the main reception desk of the university. The recurrent neural network algorithm was trained using the gathered dataset, and it performed admirably. The model attained a low loss value of 0.0167 and an accuracy of 1.0000. The results presented demonstrate the efficiency with which the RNN model performed by precisely identifying and responding to the questions that were recorded. The thesis studies the pros and cons of RNN and measures how it performs when compared to the advanced XLNet and BERT algorithms. These systems' efficiency can be gauged utilizing assessment metrics like accuracy, precision, and consistency of responses.

KEYWORDS: Chatbot; deep learning; customer service support system

1 Introduction

Technology significantly impacts business operations and daily activities. One of the promising advancements in technology is the development of chatbots, which provide an interface allowing users to interact by asking questions in simple English and receiving accurate responses from an AI-powered system [1]. Chatbots have a broad range of applications globally, especially in automating customer service, which is traditionally managed by multiple employees and is often exhausting and time-consuming [2]. The primary aim of a chatbot is to handle user inquiries efficiently, addressing questions and providing appropriate responses. To accomplish this, Artificial Intelligence (AI) plays a vital role, as it simulates human intelligence, enabling chatbots to understand human language and perform specific tasks. Researchers often use datasets like university reception datasets—and conduct experiments to evaluate the accuracy and effectiveness of AI models like XLNet and BERT in generating appropriate responses. Key evaluation metrics such as Precision, Recall, and F1 Score are used to measure how well these models capture context and produce responses that align with human-generated replies [3,4].

The application of chatbots is evident in various sectors, including the financial industry. For instance, in May 2017, Bank of America launched "Erica," a chatbot service that could perform tasks similar to Apple's Siri.

Erica provided simple text and voice responses related to transaction information, account balances, utility bill payments, fund management counseling, interest rate advice, and more [5]. It utilized machine learning and deep learning technologies to learn from clients' profiles, purchase histories, location information, and routine data, delivering precise and customized assistance [6]. This marked a significant advancement in customer service automation, showing how chatbots could offer immediate responses, provide accurate information, and reduce the workload on staff.

The evolution of chatbots can be traced back to the question posed by Alan Turing in his 1950 paper, "Computer Machinery and Intelligence", where he asked, "Can machines think?" Since then, chatbots have evolved to become more natural, knowledgeable, and mechanically sophisticated [7,8]. From a technological perspective, chatbots represent a progression of question-answering systems utilizing Natural Language Processing (NLP) [9]. The first wave of AI chatbots emerged around 2016, with social media platforms like Facebook enabling developers to create chatbots for various services, allowing users to perform daily activities within messaging platforms. Chatbots build self-learning models using computer algorithms and mathematical calculations to provide real-time responses that closely match user requests.

Despite these advancements, creating efficient chatbots remains a challenging research problem due to issues in the field of NLP and the difficulty of mimicking human speech [10]. A chatbot functions as a sophisticated software program bridging the gap between humans and bots. It serves as a crucial communication channel for businesses to provide both pre-sale and post-sale customer support. Instant responses and accurate information delivery are essential for businesses, as every client inquiry must be addressed promptly to ensure success. This highlights the importance of chatbots in reducing the workload on customer service teams and improving efficiency [11,12].

Educational institutes that struggle with old student or new student engagement can often utilize digital transformation solutions to help discover more information from their organization. For a university, the information required is often provided after some time. Customers can quickly access the suggested chatbot application, which answers redundant questions anywhere, at any time [7,13,14].

While numerous chatbot models and architectures exist, selecting the most appropriate one for specific domains, such as a university reception chatbot, requires careful consideration. Most previous research has focused on generic chatbots or specific domains, with limited studies on chatbots tailored for university reception environments. This paper aims to address this gap by focusing on the development and evaluation of a chatbot specifically designed for university reception settings. This involves analyzing its effectiveness, ensuring it provides accurate and relevant responses, and understanding its ability to handle a wide range of inquiries in a natural and context-aware manner [15]. Over the years, from simple decision trees in computerized telephone systems in the 1980s and 1990s to more sophisticated AI models today, chatbot technology has significantly evolved, yet challenges remain in creating more natural and context-aware systems [9]. Hence, this study aims to design a customer service support system (A receptionist for a university) to facilitate students with the information they need for a certain topic. The chatbot will answer the FAQs of user. It will be available 24/7 and will guide the user in their queries. The main goal of this research is to develop and evaluate a chatbot system for university reception that enhances the customer service support and communication experience within the university community. The objectives of this study are:

- To design a dataset using FAQ's and some of the receptionists' information.
- To design a DL-based model for customer service support systems.
- To design and develop a chatbot application for university reception, utilizing natural language processing, DL, and intelligent response generation techniques.

The rest of this study is organized as follows: Section 2 presents the related work. Sections 3 and 4 present the research methodology and results analysis and discussion, respectively. Finally, Section 5 concludes the study along with future work.

2 Related Work

Customer service support systems play a crucial role in enhancing user experience and satisfaction. They aim to address user inquiries, provide assistance, and offer relevant information promptly and accurately. Traditional customer service approaches, such as phone calls or email interactions, often face challenges in terms of response time and scalability. To overcome these limitations, organizations have turned to automated solutions like chatbots to streamline their customer service processes. Software agents that impersonate humans are known as chatbots. Those have built-in artificial intelligence agents that employ the processing of natural languages (NLP) for responding to requests from users. Developing the answer to the inquiry is made simpler by a prepared knowledge base [16].

In 2017, K. Chaitrali et al. Chatbot development frameworks are available, but they also make use of techniques that are either rule-based or pattern-based. Implementing rules such as If X then Y then For rule-based chatbots, which are the easiest to create, if A then B, etc., is required. As a result, if there are 100 possibilities, the developer must create 100 rules, one for every case. Such strategies are inadequate because of the volume, data diversity, and complexity [1].

In 2021, Richa Ranveera et al. The paper concentrates on developing a chatbot that would address user inquiries and offer responses. Companies typically employ a large number of people to deal with client inquiries, but this is a laborious and time-consuming task. Chatbot was developed to be used for resolving these issues. The majority of the time, text is used to store the commonly asked questions by customers, the answers, and information about the company [2].

In 2018, Fabio Clarizia et al. The creation of efficient Chatbots is one of the most difficult research topics; copying human conversation is a very tough job that entails issues with the NLP study field. It is feasible to understand what the user is writing and what their requests are, thanks to the application of NLP algorithms and approaches. The heart of the system is generally represented by this task, but there are a few issues. It is impossible to map all user requests, and the current Chatbots do not perform remarkably due to the unpredictable nature of human thought throughout a conversation [10]. In 2018 lalwani et al. Most of the time, getting all the information on a single interface is challenging without having to navigate numerous forms and windows. The chatbot for colleges aims to overcome this difficulty by providing a common and user-friendly interface to respond to inquiries from college students and lecturers [16].

In 2020, Aishwarya Gupta. We create customer service chatbots to aid businesses in having 24/7 automated responses. After studying the information and seeing how crucial it is for businesses and customers to have automatic responses. The dataset can be updated to boost the accuracy of the questions if they are unable to respond. By responding to all user inquiries, this chatbot strategy will help to increase customer happiness [9].

In 2021, Sewoong Hwang, Jonghyuk Kim. According to customer service and chatbot channels, the data demonstrate that investing in new items and maintaining existing services has a substantial impact on the rise or fall in bank earnings. This means that rather than a chatbot, Savings from housing subscriptions and new product-oriented funds are better suited for customer assistance. On the other hand, chatbots are more suited for processing services for products that already exist, such as the payment of loan interest or electric bills bank's net income [6].

In 2021, Andr'e Barbosa, Alan Godoy. Customer service personnel manually marked up data from 639,159 contacts between May 2019 and August 2020 from the entire dataset, which only comprised user communications received before an agent entered the conversation. We used an out-of-time split to separate the data into training (511,327 chats), validation (63,916 talks), and testing (63,916 chats) sets. The training set initially contained 306 different classes, but the final dataset only had 235 distinct labels after eliminating all tickets from classes with fewer than 50 samples. To fine-tune BERT on our data, because it is a very heavy model, we used a smaller subset of the dataset, with 178,578 samples for training, 19,843 for validation, and 66,141 for testing [17].

In 2022, Regin et al. The total number of different words contained in the data set determines how many nodes are present in the input and hidden layers. Whereas the number of nodes in the output equals the number of different tags used to segment the data collection. Because it doesn't require a lot of processing capacity for either training or deployment, this type of neural network is ideal for creating straightforward chatbots. We created a chatbot for a coffee business that can order coffee, deliver jokes, propose drinks, and other things. Despite being quite straightforward, this chatbot is highly adaptable, making it simple to use in any situation [18].

In 2019, Kamita et al. As the value of maintaining excellent mental health has gained widespread recognition, research into online courses on mental healthcare has picked up. Maintaining employees' mental health in the workplace has even recently become a legal requirement in Japan, which has resulted in a sudden rise in the number of people who could benefit from mental counseling, the majority of whom are not sick, but not an increase in the number of psychologists or counselors who typically provide that counseling. To address this problem, there is a strong need to give people the tools they need to be careful of their mental health on their own, and use the information on Worker mental wellness to facilitate more efficient expert collaboration [19].

In 2018, Hiremath et al. To replace a human being in the educational system's response to user inquiries, this study aims to develop an automated system. It might answer every query that the user submits. Chatbots that were already in use, including those on Facebook, WeChat, Operator, Natasha from Hike, etc., offered responses from their regional databases. Our approach, however, is to prioritize both local and online databases while also building a scalable, intuitive, and highly interactive system [20].

Notable developments in emotionally intelligent and context-aware chatbot systems have been highlighted in recent literature, particularly in service-oriented and educational contexts. A new use of AI in educational tutoring settings is demonstrated by Favero et al. (2024), who present a Socratic chatbot based on locally run Llama 2 models that successfully promotes critical thinking by posing thoughtful questions rather than offering answers [21]. When Brun et al. (2025) investigate emotion-sensitive LLM-based conversational agents, they discover that even when resolution outcomes stay the same, chatbots that can recognize and react to users' emotional tones greatly increase perceived competence and customer satisfaction [22].

The study "Empathic chatbots: A double-edged sword in customer experiences" (2025) highlights significant design trade-offs in another area [23]. It finds that while empathy in chatbots can increase social presence and satisfaction, it can also backfire when time pressure is present. ScienceDirect. According to research released in mid-2024, which supports these findings, consumers express less satisfaction and intent with chatbot interactions than with human agents unless chatbots exhibit perceived empathy, which balances assessments in both positive and negative outcome scenarios. ScienceDirect.

Practically speaking, Ed, a multilingual AI assistant introduced in early 2024 for the Los Angeles Unified School District, demonstrated how chatbots can assist students by gathering academic information and creating customized lesson plans [24]. This is an example of a scalable, curriculum-aligned use of

conversational AI in education on Wikipedia. Emotionally intelligent, multimodal, generative AI chatbots are predicted to manage almost all customer interactions by 2025, providing real-time empathy detection, proactive support, and high efficiency across platforms BrandXRAqila Media. Finally, emerging market analyses showcase a wider industry shift [25].

3 Research Methodology

This section presents the overall research design and procedure to design and develop the chatbot for the university reception. The research begins with gathering and preparing data, which is then used to train different models like RNN, XLNet, and BERT. To see how well these models perform, they are compared based on accuracy, loss, and their ability to handle queries from the university reception staff. Using supervised learning, the study relies on a labeled dataset of user questions and responses, carefully chosen to cover a range of topics about university services. The RNN model, built with TensorFlow's Keras, is particularly effective at understanding conversation flow and delivering meaningful replies. After training, the models are evaluated, and the findings are analyzed to highlight their strengths and areas for improvement. The research concludes with insights into the study's outcomes and suggestions for future work. The overall research flow is presented in Fig. 1.

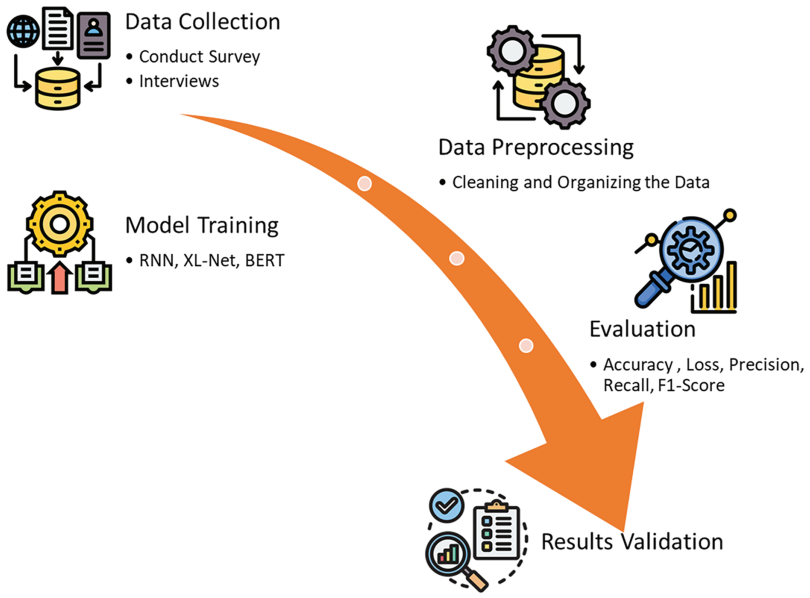

Figure 1: The proposed research flow

3.1 Data Collection

The dataset for understanding inquiries related to university reception was collected manually by a researcher person who visited various universities. This extensive data collection process involved meeting with different stakeholders, including administrative staff, students, and other relevant personnel, to gather a comprehensive range of concerns and their corresponding intents. Through direct interaction and observation, the researcher compiled a dataset of approximately 400 records, each representing a distinct user query-response pair. The data encompasses 50 attributes covering various subjects, such as general inquiries, course information, scholarships, and campus services. This approach ensured a diverse and representative dataset that accurately reflects the different types of inquiries encountered at university receptions. Table 1 presents the basic details of the dataset.

Table 1: Data description

Attribute	Description
Source of data	Actual requests from university offices and students
Number of attributes	50
Types of attributes	Text (String): Natural language queries and responses Categorical: Query labels (e.g., student query, administrative query)
Subjects covered	General Inquiries, Course Information, Scholarships, Campus Services
Number of records	400
Nature of records	Each record represents a distinct user query-response pair

3.2 Data Preprocessing

The main purpose of the data cleaning process was to make certain the dataset applied for assessment and training was accurate and free of anomalies. To do this, several crucial methods and procedures were used. Preprocessing is performed on the information that was obtained to sort it up and get it ready for additional evaluation. Tokenization and lemmatization are two text classification techniques used to normalize and classify textual data.

The steps used in preprocessing the data in this study are:

Missing Data Handling: Mean imputation and forward-fill/back-fill were the two primary approaches we employed to address missing data. The mean of the available data was used to fill in any missing values for numerical variables like the timing of responses. We utilized forward-fill or back-fill methods to solve text-based inquiries and answers, filling in the missing information with the previous or subsequent value to maintain meaning. The isna() function, which can be used to find missing data in pandas DataFrames, and fillna(), which can be used for assuming missing values, were used to locate missing values.

Eliminating Duplicates: To avoid bias and duplication in the training dataset, identical query-response sequences were detected and eliminated. This made it easier to keep a broad and objective dataset. Based on query-response combinations, we used the dropduplicates() method to find and eradicate duplicate records.

Outlier Handling: The training of models can be greatly affected by the possibility of outliers in attributes like inquiry duration or time to response. To rectify this, we detected and monitored outliers. Data points that considerably varied from the mean were reduced using Python modules such as NumPy and pandas.

Tokenization: In this phase of the process, the user's question is broken down into distinct keywords or tokens [26]. The data tokens for the example question can be seen in Fig. 2.

Lemmatization: Likewise, as stemming, lemmatization additionally accommodates the word's meaning and context into consideration. Words are simplified to their lemma, or dictionary-based form. The word "tell" has been simplified by lemmatization to its simplest or dictionary form, which is "tell" [27]. To maintain the word's conceptual integrity, this strategy takes the word's meaning and historical context into account when making decisions. Fig. 3 is a representation.

3.3 Testing and Training of Data

The datasets used to carry out our research project "Customer Service Support System (Chatbot for University Reception)" are discussed in this part, along with the training and testing standards for those datasets. We used the hold-out approach to divide the dataset into training and testing subsets, ensuring accurate model analysis of performance. This division's main goal is to train the models on enough data while also providing a distinct set of scrutiny queries that are not visible to the public. There were two sections to the data.

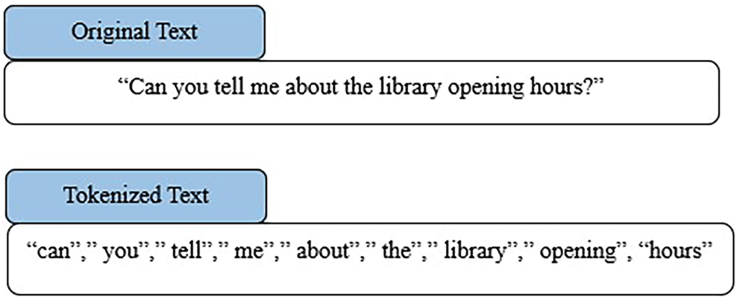

Figure 2: Example of tokenization

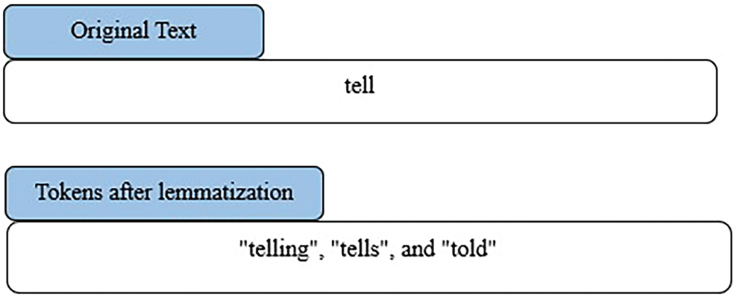

Figure 3: Example of lemmatization

80% of the dataset was set aside for training. The RNN, XLNet, and BERT models were trained using this training data to figure out interactions among queries and their associated intentions for university reception. To reduce the learning loss and improve the model's prediction ability, the internal parameters of the model have been optimized during the training phase.

The trained models were tested using the final 20% of the dataset. Unseen questions from the training phase are included in the testing data. We can evaluate the models' capacity to generalize and reliably categorize queries in a real-world context by comparing them to results from this independent testing dataset. Presenting the results of the testing queries to the models that were trained and comparing the anticipated intents with the actual truth labels comprised the testing criterion. This comparison enables us to evaluate each model's classification accuracy, precision, recall, and F1 score. The validation procedure's assessment metrics provide us with information about how well-trained the models are at responding to inquiries from university receptionists in everyday life. We ensure a thorough assessment of the RNN, XLNet, and BERT models in our chatbot system by following certain training and testing standards. While the testing step evaluates the models' adaptability skills and performance on hypothetical statements, the training process enables the models to acquire knowledge from training data. This technique makes it possible to assess

the models' precision and applicability for efficiently responding to user inquiries in the context of the university reception.

3.4 Techniques Employed

We utilized a variety of strategies in our research project, "Customer Service Support System (Chatbot for University Reception)," to examine the efficiency of RNN, XLNet, and BERT models in handling questions at a university reception. RNNs with Long Short-Term Memory architecture, in particular, were used in the code. Challenges like maintaining long-term dependencies in sequential data are addressed by this LSTM design [28]. The algorithm can forecast responses for fresh user messages because it was trained on a sample of intent-response pairings. The mentioned methodologies also include analysis of literature and LSTM networks. While LSTM networks use deep learning for natural language creation, bibliometric analysis measures papers quantitatively. For training purposes, this project uses more than a million Twitter interactions. LSTM is a useful tool for applications like language modeling and speech recognition because it excels at comprehending and remembering long-term patterns in sequential data. This study intends to show how effective these methods are at creating a user-friendly chatbot system that is suited to the needs of university reception.

3.4.1 Recurrent Neural Network

The RNN architecture can be shown as a sequence of interconnected nodes, representing different layers and their connections. Each node in the network of connected nodes that makes up the RNN architecture can be thought of as a layer in the model. A recurrent layer analyzes sequences by updating internal states, an output layer produces predictions, and an embedding layer encrypts the input text. Sequential patterns can be captured by the model thanks to information flowing across the connections between nodes [29]. Traditional RNNs, on the other hand, have trouble with long-term dependencies because of vanishing gradients. LSTM networks, which combine memory cells and gating mechanisms to improve the model's capacity to learn and retain pertinent information across longer sequences, were created as a result of this issue. The model now performs substantially better when handling sequential data, such as natural language, due to this architectural improvement [30].

3.4.2 Transformer-XL Network

XLNet is a popular transformer-based machine learning approach used for natural language processing tasks. XLNet incorporates the latest developments in NLP and addresses issues while introducing additional methods for language modeling. One notable feature is its auto-regressive language model, which facilitates joint predictions on a sequence of tokens, aiming to capture collective changes in word tokens within a sentence. The XLNet model consists of two phases: pre-training and fine-tuning. The pre-training phase is the primary focus of XLNet and introduces a new goal known as Permutation Language Modeling. XLNet architecture for analyzing movie reviews, several steps are involved. Firstly, the input text is divided into sub-words, and then positional embeddings are added to preserve the sequence order. To eliminate the left-to-right context dependency, the input sequence is randomly permuted. The permuted sequence is then passed through transformer layers, which consist of self-attention and feed-forward sub-layers. An inverse permutation layer is used to restore the original sequence order. Finally, similar to traditional language models, the hidden states of the transformer layers are utilized to predict the next token in the sequence. XLNet offers advantages such as bidirectional context learning, a two-stream self-attention architecture for understanding content and position, and training on a larger dataset to enhance language comprehension. In this experiment, the hyperparameters that were tuned include the number of epochs, training batch size,

maximum sequence length, learning rate, and weight decay [31]. The objective is to find the best values for these hyper-parameters to achieve optimal performance in the given task.

3.4.3 Bidirectional Encoder Representations from Transformers

BERT is a pre-trained language model that has revolutionized NLP tasks. It employs a bidirectional transformer to capture the contextual meaning of words, leading to improved accuracy in natural language understanding. BERT has achieved impressive results in question-answering, sentiment analysis, and text classification [32]. It can be fine-tuned on specific tasks using labeled data. BERT is widely used in real-world applications such as chatbots, search engines, and language translation. Each word in the input sequence is converted into a fixed-length vector representation, which is then processed by 12 self-attention layers. The transformer encoder generates a vector representation sequence for the input sentence. BERT's framework involves pre-training on a large unlabeled corpus and fine-tuning on labeled data. It offers the advantages of bidirectional learning, efficient size, and open-source availability [33]. The experiment tunes hyperparameters like epochs, batch size, sequence length, learning rate, and weight decay for optimal performance.

3.5 Assessment Criteria

To evaluate the proposed model and benchmark it against other techniques, several assessment criteria were employed. These criteria aimed to evaluate the performance and accuracy of the models. The following assessment metrics were used for evaluation:

1. Accuracy:

Accuracy measures the percentage of proper guesses classified as responses by the chatbot over the total number of responses. It indicates how well the chatbot's predictions match the actual responses. The equation for accuracy is:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{1}$$

2. F-Measure:

F-measure provides a balanced assessment of the Chatbot's performance by combining precision and recall into one score. Recall is the percentage of properly anticipated positive responses to the total number of positively predicted responses, whereas precision is the percentage of positively predicted responses to the total number of positively actualized responses. The formula for F-measure is:

$$F - measure = 2 * \frac{Precisopn \times Recall}{Precision + Recall} \tag{2}$$

3. Precision

Precision is a metric for how accurately a model predicts the good outcomes. It is the ratio of successfully predicted positive cases, or true positives, to all positive predictions, including both true and false positives. The equation for precision is:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{3}$$

where:

- True Positives: The number of instances that are correctly predicted as positive.

- False Positives: The number of instances that are incorrectly predicted as positive.

Precision ranges from 0 to 1, where a higher value indicates a lower rate of false positive predictions. High precision means that the model's positive predictions are mostly accurate.

4. Recall

Recall is the proportion of correctly predicted positive observations to all genuine classes.

Observations:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{4}$$

This equation calculates the proportion of actual positive instances that were correctly identified by a classification model.

5. Percentage Difference

The percentage difference formula calculates the relative difference between two values. It involves finding the absolute difference between these values and then normalizing it by the average of the two values. The result is multiplied by 100 to express the difference as a percentage. This metric provides a way to understand how much value A differs from value B in terms of percentage:

$$Percentage\ Difference = \frac{|Value\ A - Value\ B|}{\left[\frac{(Value\ A + Value\ B)}{2}\right]} \times 100 \tag{5}$$

4 Results Analysis and Discussion

This chapter demonstrates the experimental findings made possible by applying the proposed methodology in the context of our project. The primary objective of this chapter is to analyze and evaluate the performance of the developed models or techniques in addressing the research questions and achieving the desired outcomes. Next, the chapter focuses on the techniques employed in the research project, with a particular emphasis on comparing the accuracy and error rates of different models such as RNN, XLNet, and BERT. However, due to the limited size of the dataset, the RNN model was chosen for this project, as it demonstrated better performance on smaller datasets. Even though BERT performs better than the previous baselines for each NLP task and is simple to apply, it has limitations if the task corpus is overly narrowly focused on a certain topic. The model is assisted in training contextualized representations of words that do not appear in broad corpora (such as English Wikipedia) by post-training on domain-specific corpora (such as the Ubuntu Corpus) [32]. Furthermore, the chapter includes visual representations such as tables, figures, or charts to illustrate the experimental results effectively. These visual aids help in presenting the findings and comparing the performance of different models or techniques. Our experimental findings demonstrate that bidirectional LSTM models can perform much better on a limited dataset than a BERT model and that training these straightforward models takes a lot less time than fine-tuning their more complex equivalents. We conclude that a model's performance is reliant on the job and the data; hence, these factors should be taken into account before picking a model, rather than just going with the most often used model [34].

4.1 Experimental Results

Each row represents a different algorithm, and the columns display specific information about their parameters. "Total Params" refers to the total number of parameters within each model, encompassing all parameters, whether trainable or not. "Trainable Params" indicates the number of parameters that are adjusted during the training process. These are the ones that the model learns from the data. "Non-trainable

Params" represents the number of parameters that remain fixed during training, often associated with pre-trained components. For all three algorithms, RNN, BERT, and XLNET, the total number of parameters is the same, being 17,309. The "Trainable Params" value is identical to the total parameters in each case, suggesting that all the parameters are trainable. "Non-trainable Params" is consistently zero, implying that there are no fixed or non-trainable parameters in any of the models. This comparison of model parameters helps us understand the complexity and composition of each algorithm's architecture, aiding in evaluating their computational requirements and potential performance. Table 2 shows the results obtained from the experiments conducted using the algorithms.

Table 2: Model parameters comparison

Algorithm	Total params	Trainable params	Non-trainable params
RNN	17,309	17,309	0
BERT	17,309	17,309	0
XLNET	17,309	17,309	0

The results shown in Table 3 show that the RNN algorithm achieved a high accuracy of 1.0000 and a validation accuracy of 0.0714. The BERT algorithm, on the other hand, yielded lower accuracy and validation accuracy values. The XLNET algorithm exhibited moderate performance with an accuracy of 0.12. These outcomes provide insights into the performance of the employed algorithms and their ability to correctly classify instances. The simplest LSTM design that we tried outperformed the best for this dataset after we experimented with a variety of LSTM topologies. LSTM statistically considerably outperformed XLnet, BERT, and test data accuracy when compared to validation data and test data, respectively. Additionally, the experimental findings demonstrated that XLnet overfits more than the LSTM architecture for smaller datasets, BERT [34]. The analysis of these results helps evaluate the effectiveness of each algorithm and aids in selecting the most suitable approach for the given task. Figs. 4–6 show the accuracy and loss metrics for each model. It is important to note that accuracy alone may not provide a complete understanding of the model's performance, and other evaluation metrics like precision, recall, and F1-score should also be considered for a comprehensive analysis, which is in Table 3.

Fig. 6: XLNet Accuracy and Loss Table 4. The experiments were performed over multiple epochs, and the classification metrics such as accuracy, loss, precision, recall, and F1-score were recorded. Table 4 summarizes the performance of three algorithms, RNN, BERT, and XLNET, in terms of precision, recall, and F1-Score. Precision measures accurate positive predictions, recall gauges the correct identification of actual positives, and the F1-Score balances both aspects. Higher values signify better performance. The RNN shows the highest precision but lower recall, while XLNET has the highest recall but moderate precision. BERT performs relatively lower on both precision and recall. This comparison aids in understanding each algorithm's effectiveness in the classification task, considering false positives and false negatives.

Table 3: Performance metrics of models

Algorithm	Epoch	Accuracy	Loss
RNN	550	1.0000	0.0167
BERT	550	0.0476	3.5472
XLNET	10	0.12	5

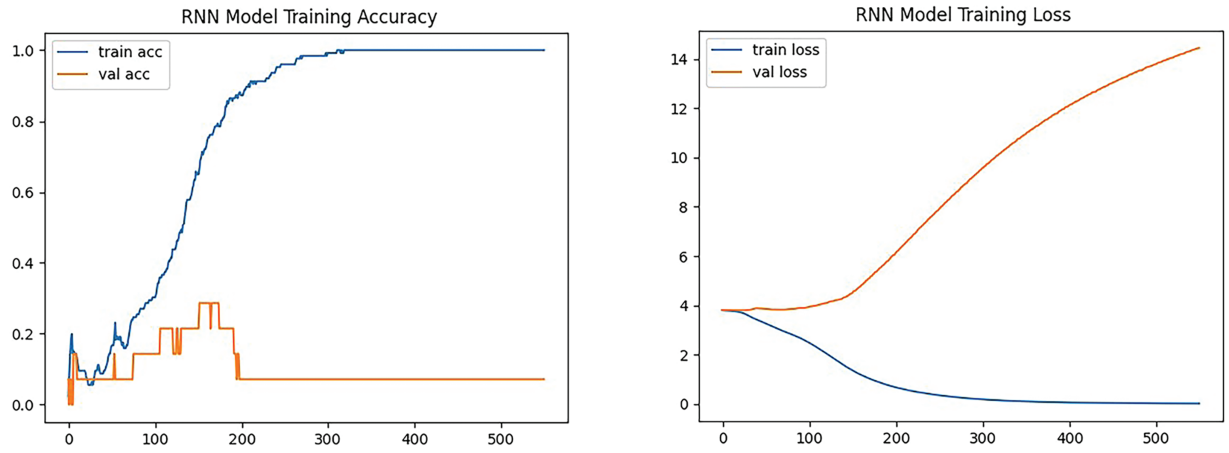

Figure 4: RNN accuracy and loss

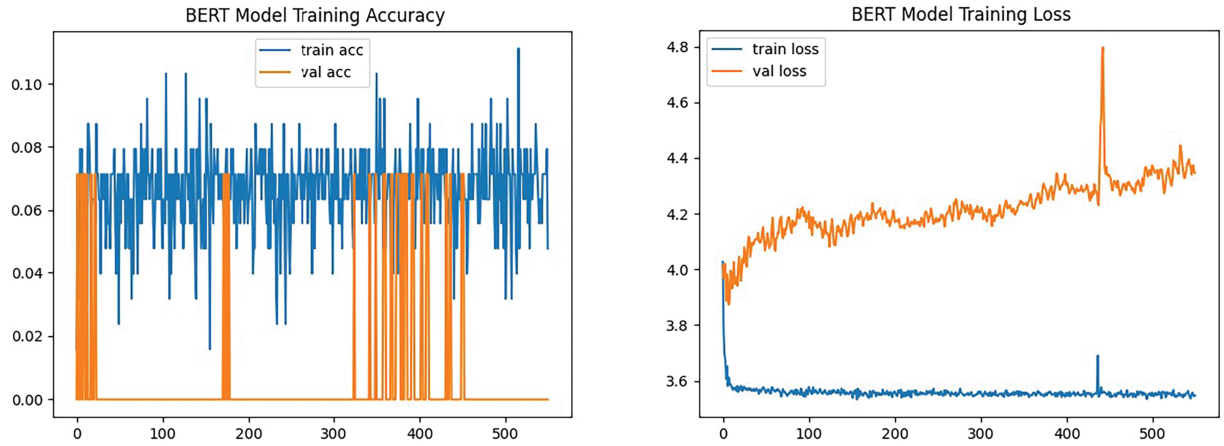

Figure 5: BERT accuracy and loss

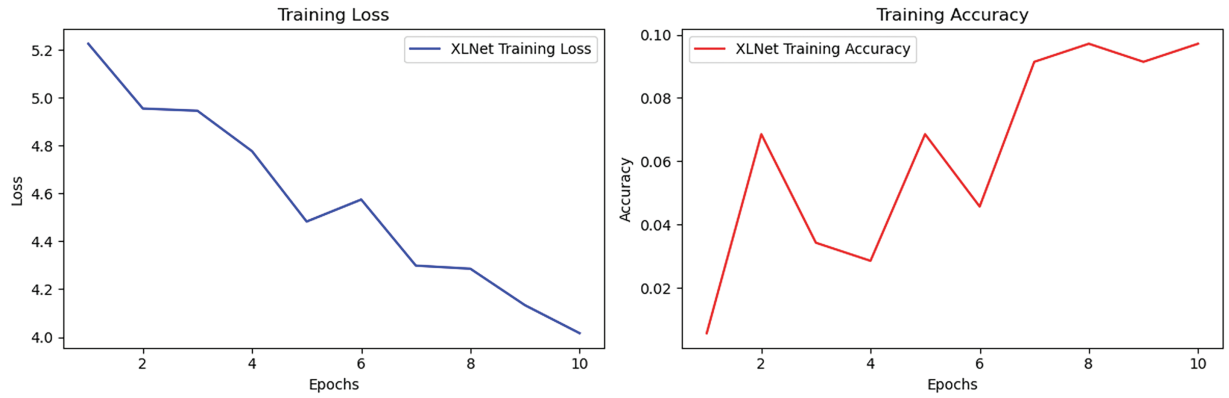

Figure 6: XLNet accuracy and loss

Table 4: Results of models

Algorithm	Precision	Recall	F1-Score
RNN	0.2771	0.2285	0.21904
BERT	0.003265	0.05714	0.006177
XLNET	0.06021	0.10285	0.04031

Figs. 7 and 8 show how to start a discussion with the CUSIT Chatbot. The user-friendly design offers a platform for users to connect with the chatbot. The text input field's users can just start typing their inquiries or questions there. The user's first interaction is shown in the figure by entering "Hello" into the chat box. Users can effortlessly start conversations and look for information due to this interaction, which serves as an introduction to the chatbot's conversational experience [35]. The chatbot processes the user's text input to process the question and produce an appropriate response, starting a dynamic and interactive conversation.

Table 5 provides a comparison of percentage differences in various metrics between different models (RNN, BERT, and XLNET). The metrics include Accuracy, Loss, Precision, Recall, and F1 Score. The "Models" column indicates the comparison being made (e.g., RNN vs. BERT, RNN vs. XLNET), while the "Percentage Difference" column displays the calculated percentage difference for each metric between the two models being compared. The percentages illustrate how much one model's metric differs from the other models' metric.

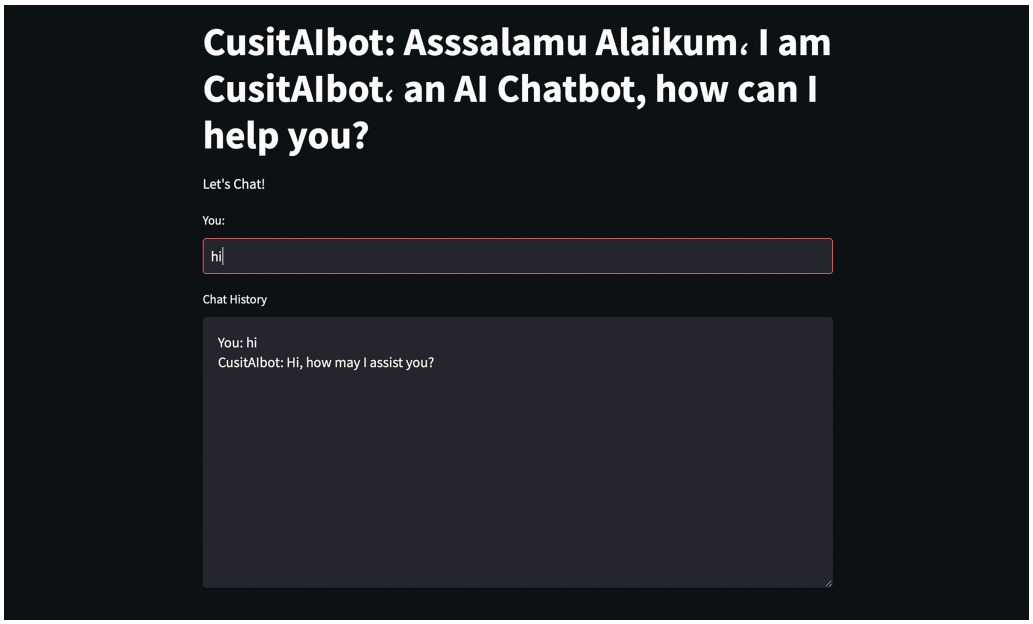

Figure 7: User interaction with CUSIT chatbot

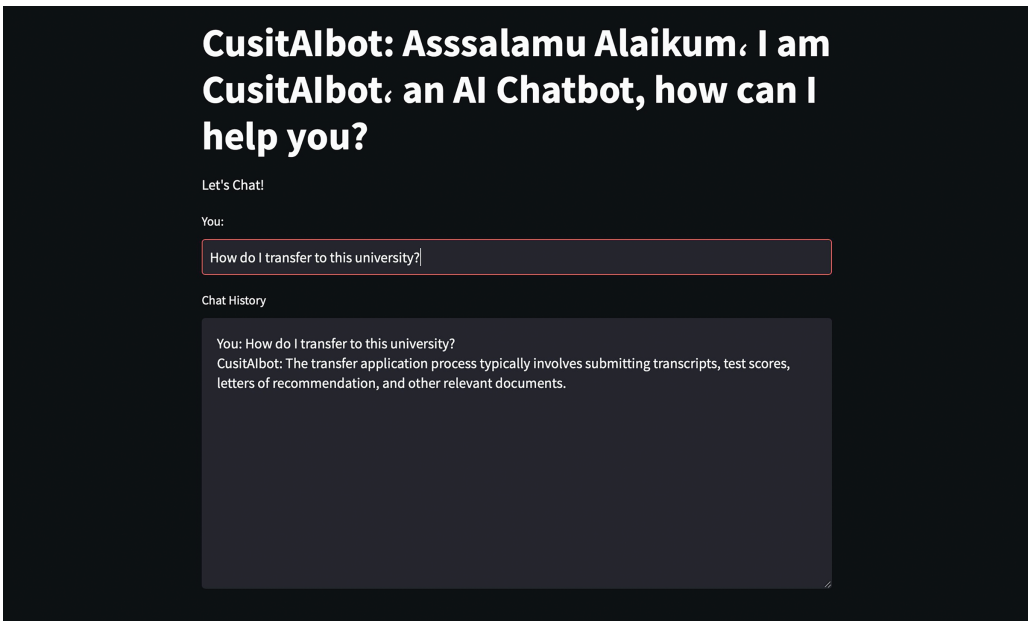

Figure 8: Test result of ChatBot

Table 5: Percentage difference comparison

Measurements	Models	Percentage difference
Accuracy	RNN VS BERT	181.825%
	RNN VS XLNET	157.143%
Loss	RNN VS BERT	−198.126%
	RNN VS XLNET	−198.668%
Precision	RNN VS BERT	195.342%
	RNN VS XLNET	128.6%
Recall	RNN VS BERT	119.983%
	RNN VS XLNET	75.8413%
F1 Score	RNN VS BERT	189.029%
	RNN VS XLNET	137.829%

We used the percentage difference to compare pairs of metric values across models in a scale-free way:

$$Percentage\ Difference = \frac{|A - B|}{\frac{A+B}{2}} \times 100 \tag{6}$$

For example, comparing accuracy of RNN (A = 1.000) and BERT (B = 0.0476):

$|A - B| = |1.000 - 0.0476| = 0.9524$, $\frac{|A-B|}{2} = \frac{1.000 + 0.0476}{2} = 0.5238$, so $\frac{0.9524}{0.5238} \times 100 = 181.83\%$

This indicates the two accuracies differ by ~181.83% relative to their average, i.e., a very large gap favoring the RNN on this dataset.

The comparison of different metrics between the RNN model and both BERT and XLNet models reveals substantial variations in performance. Regarding accuracy, the RNN model significantly outperforms

both BERT and XLNet models by 181.825% and 157.143%, respectively. In terms of loss, the RNN model exhibits negative values, indicating a substantial decrease in loss compared to BERT and XLNet models by −198.126% and −198.668%. The precision of the RNN model stands out as it surpasses BERT and XLNet models by 195.342% and 128.6%. In terms of recall, the RNN model demonstrates superiority with a 119.983% higher recall than BERT and 75.8413% higher recall than XLNet. Lastly, the F1 score of the RNN model is considerably higher, showing improvements of 189.029% and 137.829% over the BERT and XLNet models, respectively. These differences underscore the RNN model's superior performance across various evaluation metrics. These percentage difference values provide insights into how the RNN model performs relative to the BERT and XLNet models across different evaluation metrics. Positive values indicate that the RNN model outperforms the other models, while negative values indicate that it underperforms in terms of the specific metric.

4.2 Discussion

The outcomes of this study shed light on the results obtained from our project. Our research focused on analyzing a specific dataset, which contains various attributes relevant to our problem domain. One crucial aspect of the dataset was the varying percentage of effective and non-effective instances, which played a significant role in our analysis. By applying different machine learning algorithms to the dataset, we aimed to evaluate their performance and understand how they handle the given attributes. The results of this study highlighted the reliability with which each algorithm categorized instances based on the data that was available. They availed the use of RNN-based models, namely a bidirectional long short-term memory and a vanilla RNN. The RNN-based encoder anticipates the probability score of a particular reply being the next statement in a given conversation context [36] by using sequential representations that represent a dialog context and a response. We discovered that the accuracy of the techniques varied across datasets during our investigation. The distinction was related mostly to the unique qualities contained in each dataset, as well as the proportion of effective and ineffective circumstances. Determining the dataset characteristics was essential to comprehending algorithm performance.

In this study, the evaluation metrics were carefully selected to align with the research goal of validating the chatbot's ability to improve customer service and communication within the university community. Accuracy was used to assess the overall correctness of the chatbot's responses, reflecting its reliability in handling user queries. Precision measured the proportion of correctly generated responses among all responses, which is particularly important in minimizing irrelevant or misleading answers. Recall evaluated the chatbot's ability to retrieve all relevant responses, ensuring that no essential information was omitted in user interactions. The F-measure, as the harmonic mean of precision and recall, was employed to provide a balanced view of the system's ability to maintain both correctness and completeness. These metrics were preferred over alternatives such as response time or user satisfaction ratings at this stage, as they offer objective, quantifiable measures directly linked to the dataset and the intended function of the chatbot system.

On the dataset, we discovered that the LSMT RNN performed better in terms of accuracy, precision, recall, and f-measure. It indicates the algorithm that outperformed the others. However, while models based on transformers show potential in producing classification results with relatively high levels of accuracy, they are likely to fall short in terms of integrity and their interpretability when compared to more conventional machine learning algorithms, such as random forest, because transformer models have many more parameters and an advanced architecture [36]. These findings highlight the significance of selecting the best algorithm depending on the properties of the dataset and the intended assessment measures. It

also emphasizes the need to have a good grasp of the dataset, particularly the distribution of cases, to make appropriate decisions when selecting a technique for the task of classification.

The findings of this study align with recent research on the application of AI-driven chatbots in educational and customer service contexts. For example, Nam et al. (2024) introduced NOVI, a chatbot system leveraging GPT-based models to support university freshmen, highlighting the role of conversational agents in enhancing communication and information access within academic institutions. Similarly, Thway et al. (2024) demonstrated the effectiveness of a Retrieval-Augmented Generation (RAG) chatbot in improving student engagement and learning experiences, reinforcing the importance of accurate and context-aware responses. In line with these studies, our chatbot system demonstrates that well-defined evaluation metrics such as accuracy, precision, recall, F-measure, and percentage difference can effectively validate system reliability, thereby supporting improved customer service at the university reception desk. This confirms that our work not only contributes practically to enhancing communication in the university community but also extends the growing body of research on the integration of intelligent chatbots in higher education service delivery.

Accuracy was considered acceptable when exceeding 80%, reflecting a reliable level of correct predictions relative to the total responses. Precision indicates the proportion of correctly predicted positive responses, and values above 0.75 were regarded as satisfactory for minimizing false positives. Recall was deemed sufficient when above 0.70, ensuring that the chatbot successfully retrieved a majority of relevant responses without excessive omission. The F-measure, representing the harmonic mean of precision and recall, was used to balance these two aspects, with values approaching or exceeding 0.75 reflecting good performance. Finally, the percentage difference was used to assess the deviation between predicted and actual values, where lower values signified higher validity of the chatbot outputs. Together, these thresholds confirm the validity and reliability of the chatbot's responses within the university reception context.

This study provides useful information regarding the outcomes of our efforts. It emphasizes the need for good algorithm selection to obtain accurate classification results, as well as the impact of dataset properties and instance distributions on algorithm performance. They suggested that small amounts of data might not be the ideal choice for fine-tuning the complete BERT layers [32].

5 Conclusion and Future Work

In conclusion, this research aimed to develop a university reception chatbot and compare the performance of RNN, BERT, and XLNet models. The chatbot assists users in obtaining information about university services. RNN demonstrated context understanding, while BERT excelled in semantic comprehension, and XLNet handled bidirectional context effectively. The choice of model should align with the query nature, training data, and accuracy needs. The study enhances AI-driven conversational systems in education and aids universities in implementing efficient chatbots for improved user experiences. To further improve accuracy and satisfaction with users, future research can extend datasets, add advanced NLP approaches, and improve chatbot functionality. A chatbot system's objective is to imitate human communication. Its structure simulates online communication utilizing natural language between a human and a machine by combining a computational algorithm and a language model. The development of the university reception chatbot and the comparison of RNN, BERT, and XLNet models contribute to the advancement of AI-driven conversational systems in the educational domain. The findings of this study can assist universities in implementing efficient and reliable chatbot solutions to enhance user experiences and streamline information retrieval processes. The university reception chatbot's development and comparison reveal potential directions for improvement. Using advanced methods like semantic parsing and pre-trained language models can enhance natural language comprehension. Increasing topic expertise through FAQs, scholarly sources, and databases

can improve response accuracy. Multi-modal capabilities, user feedback systems, and reinforcement learning can enhance the chatbot's relevance and utility. Integrating the chatbot with existing university systems, such as student databases, can increase its relevance and utility. Addressing ethical considerations like privacy, security, fairness, and bias is crucial for responsible and user-centric operations. These advancements will lead to more intelligent, efficient, and supportive chatbot systems in university settings.

Acknowledgement: The authors would like to express their sincere gratitude to the Department of Computer Science, City University of Science and Technology, Peshawar, and the Department of Computer Science, University of Engineering and Technology, Mardan, for their continuous support and provision of resources that facilitated the successful completion of this research. The authors also extend their appreciation to colleagues and peers whose constructive feedback and encouragement contributed to improving the quality of this work.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Bilal Khan, Muhammad Adeen Jamal; data collection: Bilal Khan, Sameed Ur Rehman; analysis and interpretation of results: Muhammad Adeen Jamal, Wahab Khan; draft manuscript preparation: Bilal Khan, Wahab Khan. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data utilized in this study were collected from university staff, students, and management, and will be made available upon reasonable request.

Ethics Approval: This study was conducted in accordance with ethical research standards. Participation was voluntary, and respondents were assured that their information would remain confidential and would be used solely for research purposes. No sensitive personal data was disclosed, and the study involved minimal risk to participants. As the research was limited to service evaluation and user interaction with the chatbot system, formal institutional ethical approval was not required.

Informed Consent: Informed consent was obtained from all participants, including university staff, students, and management, prior to data collection.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Kulkarni CS, Bhavsar AU, Pingale SR, Kumbhar SS. BANK CHAT BOT-an intelligent assistant system using NLP and machine learning. Int Res J Eng Technol. 2017;4(5):2374–7.
2. Ranveera R, Kesharwani A, Kumari S. Customer support chatbot using natural language processing. Int J Innov Eng Res Technol. 2021;8(06):237–41.
3. Joshi M, Pangave VV, Tamhan P, Lule A, Mhatre A, Puranik BS. Automated chatbots for improved user services in university libraries. Libr Prog Libr Sci Inf Technol Comput. 2024;44(1):176–93.
4. Islam M, Warsito B, Nurhayati O. AI-driven chatbot implementation for enhancing customer service in higher education: a case study from universitas Negeri Semarang. J Theor Appl Inf Technol. 2024;102(14):5690–701.
5. Yu S, Chen Y, Zaidi H. AVA: a financial service chatbot based on deep bidirectional transformers. Front Appl Math Stat. 2021;7:604842. doi:10.3389/fams.2021.604842.
6. Hwang S, Kim J. Toward a chatbot for financial sustainability. Sustainability. 2021;13(6):3173. doi:10.3390/su13063173.
7. Davar NF, Dewan MAA, Zhang X. AI chatbots in education: challenges and opportunities. Information. 2025;16(3):235. doi:10.3390/info16030235.
8. Zhang F, Liu X, Wu W, Zhu S. Evolution of chatbots in nursing education: narrative review. JMIR Med Educ. 2024;10(1):e54987.

9. Gupta A, Hathwar D, Vijayakumar A. Introduction to AI chatbots. Int J Eng Res Technol. 2020;9(7):255–8.
10. Clarizia F, Colace F, Lombardi M, Pascale F, Santaniello D. Chatbot: an education support system for student. In: Cyberspace Safety and Security: 10th International Symposium, CSS 2018; 2018 Oct 29–31. Amalfi, Italy. p. 291–302.
11. Oqaidi K, Aouhassi S, Mansouri K. Are chatbots the future of higher education? Unveiling their impact, challenges, and prospects. In: 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET); 2024 May 16–17; FEZ, Morocco. p. 1–6.
12. Naik I, Naik D, Naik N. Investigating the benefits and barriers of using AI chatbots in education. In: The International Conference on Computing, Communication, Cybersecurity & AI. Berlin/Heidelberg, Germany: Springer; 2024. p. 510–24.
13. Aboelmaged M, Bani-Melhem S, Ahmad Al-Hawari M, Ahmad I. Conversational AI chatbots in library research: an integrative review and future research agenda. J Librariansh Inf Sci. 2025;57(2):331–47. doi:10.1177/09610006231224440.
14. Laun M, Wolff F. Chatbots in education: hype or help? A meta-analysis. Learn Individ Differ. 2025;119:102646. doi:10.1016/j.lindif.2025.102646.
15. Khan MM. Development of an e-commerce sales chatbot. In: 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET); 2020 Dec 14–16; Charlotte, NC, USA. p. 173–6.
16. Lalwani T, Bhalotia S, Pal A, Rathod V, Bisen S. Implementation of a chatbot system using AI and NLP. Int J Innov Res Comput Sci Technol. 2018;6(3):26–30.
17. Barbosa A, Godoy A. Augmenting customer support with an NLP-based receptionist. arXiv:211201959. 2021.
18. Regin R, Rajest SS, Shynu T. An automated conversation system using natural language processing (NLP) chatbot in python. Cent Asian J Med Nat Sci. 2022;3(4):314–36.
19. Kamita T, Ito T, Matsumoto A, Munakata T, Inoue T. A chatbot system for mental healthcare based on SAT counseling method. Mob Inf Syst. 2019;2019(1):9517321. doi:10.1155/2019/9517321.
20. Hiremath G, Hajare A, Bhosale P, Nanaware R, Wagh KS. Chatbot for education system. Int J Adv Res Ideas Innov Technol. 2018;4(3):37–43.
21. Favero L, Pérez-Ortiz JA, Käser T, Oliver N. Enhancing critical thinking in education by means of a Socratic chatbot. In: International Workshop on AI in Education and Educational Research. Berlin/Heidelberg, Germany: Springer; 2024. p. 17–32. doi:10.1007/978-3-031-93409-4_2.
22. Brun A, Liu R, Shukla A, Watson F, Gratch J. Exploring emotion-sensitive LLM-based conversational AI. arXiv:2502.08920. 2025.
23. Juquelier A, Poncin I, Hazée S. Empathic chatbots: a double-edged sword in customer experiences. J Bus Res. 2025;188:115074. doi:10.1016/j.jbusres.2024.115074.
24. Markovitch DG, Stough RA, Huang D. Consumer reactions to chatbot versus human service: an investigation in the role of outcome valence and perceived empathy. J Retail Consum Serv. 2024;79:103847. doi:10.1016/j.jretconser.2024.103847.
25. Sharma A, Sharma P, Gaur R. Artificial Intelligence (AI) and the future of marketing trends: challenges and opportunities. Artif Intell Peace Justice Strong Inst. 2025;28:23–46. doi:10.4018/979-8-3693-9395-6.ch002.
26. Gkinko L, Elbanna A. The appropriation of conversational AI in the workplace: a taxonomy of AI chatbot users. Int J Inf Manage. 2023;69:102568. doi:10.1016/j.ijinfomgt.2022.102568.
27. Camacho-Collados J, Pilehvar MT. On the role of text preprocessing in neural network architectures: an evaluation study on text categorization and sentiment analysis. arXiv:1707.01780. 2017.
28. Kasthuri E, Balaji S. Natural language processing and deep learning chatbot using long short term memory algorithm. Mater Today Proc. 2023;81:690–3. doi:10.1016/j.matpr.2021.04.154.
29. Anki P, Bustamam A, Al-Ash HS, Sarwinda D. Intelligent chatbot adapted from question and answer system using RNN-LSTM model. In: 2020 2nd International Conference on Science & Technology (2020 2nd ICoST); 2020 Nov 28; Yogyakarta, Indonesia.
30. Dhyani M, Kumar R. An intelligent Chatbot using deep learning with Bidirectional RNN and attention model. Mater Today Proc. 2021;34:817–24. doi:10.1016/j.matpr.2020.05.450.

31. Salma TD, Saptawati GAP, Rusmawati Y. Text classification using XLNet with infomap automatic labeling process. In: 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA); 2021 Sep 29–30; Bandung, Indonesia.
32. Whang T, Lee D, Lee C, Yang K, Oh D, Lim H. An effective domain adaptive post-training method for bert in response selection. arXiv:1908.04812. 2019.
33. Cortiz D. Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra. arXiv:2104.02041. 2021.
34. Ezen-Can A. A comparison of LSTM and BERT for small corpus. arXiv:2009.05451. 2020.
35. Rosruen N, Samanchuen T. Chatbot utilization for medical consultant system. In: 2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON); 2018 Dec 12–14; Bangkok, Thailand. p. 1–5.
36. Terechshenko Z, Linder F, Padmakumar V, Liu M, Nagler J, Tucker JA, et al. A comparison of methods in political science text classification: transfer learning language models for politics. [cited 2025 Sep 1]. Available from: https://ssrn.com/abstract=3724644.