

Article

EduChat: An AI-Based Chatbot for University-Related Information Using a Hybrid Approach

Hoa Dinh  and Thien Khai Tran * 

Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages and Information Technology, Ho Chi Minh City 700000, Vietnam; hoadm@huflit.edu.vn

* Correspondence: thientk@huflit.edu.vn

Abstract: The digital transformation has created an environment that fosters the development of effective chatbots. Through the fusion of artificial intelligence and data, these chatbots have the capability to provide automated services, optimize customer experiences, and reduce workloads for employees. These chatbots can offer 24/7 support, answer questions, perform transactions, and provide rapid information, contributing significantly to the sustainable development processes of businesses and organizations. ChatGPT has already been applied in various fields. However, to ensure that there is a chatbot providing accurate and useful information in a narrow domain, it is necessary to build, train, and fine-tune the model based on specific data. In this paper, we introduce EduChat, a chatbot system for university-related questions. EduChat is an effective artificial intelligence application designed by combining rule-based methods, an innovative improved random forest machine learning approach, and ChatGPT to automatically answer common questions related to universities, academic programs, admission procedures, student life, and other related topics. This chatbot system helps provide quick and easy information to users, thereby reducing the time spent searching for information directly from source documents or contacting support staff. The experiments have yielded positive results.

Keywords: engineering; chatbot engineering; AI-based chatbot; EduChat; improved random forest; University Consulting



Citation: Dinh, H.; Tran, T.K.

EduChat: An AI-Based Chatbot for University-Related Information Using a Hybrid Approach. *Appl. Sci.* **2023**, *13*, 12446. <https://doi.org/10.3390/app132212446>

Academic Editor: Alexander Barkalov

Received: 12 October 2023

Revised: 15 November 2023

Accepted: 16 November 2023

Published: 17 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

“Chatbot”, an abbreviation for “chat robot”, is an artificial intelligence (AI)-based conversational entity designed to simulate human conversation. Chatbots have emerged as an innovative technology, transforming how humans interact with machines and services. This technology, driven by advancements in natural language processing (NLP) and machine learning, enables smooth and efficient communication between users and digital systems. Chatbots find applications in various diverse fields, from customer support, e-commerce, healthcare, to education, thanks to their real-time responsiveness, personalization capabilities, and task automation.

In recent years, one of the notable advancements in chatbot research is the work by Brown et al. (2020) [1]. This paper introduces GPT-3 (Generative Pre-trained Transformer 3), a language model with a staggering 175 billion parameters. GPT-3 demonstrates remarkable language understanding and generation abilities by producing human-like responses. Prior to that, the work of Zhang et al. (2019) [2] introduced DialoGPT, a variant of the GPT-2 model fine-tuned for generating responses in conversations. By fine-tuning on conversational data, DialoGPT enhances the fluency and relevance of the generated responses. Furthermore, it is also worth mentioning a chatbot in the open research domain created by Adiwardana et al. (2020), named Meena [3], which utilizes adversarial training methods to generate interesting and contextually relevant responses. This research highlights the importance of creating meaningful conversations. Particularly in [4], Reddy et al. (2021)

introduced the DALL-E system—an image generation system based on text descriptions. This opens up the possibility of expanding the communication capabilities of chatbots beyond text, allowing for the display and comprehension of images.

These research advancements demonstrate the unique evolution of chatbots, pushing the boundaries of what AI-driven entities can achieve. From comprehensive language understanding to image generation and maintaining meaningful conversations, these breakthroughs highlight the potential of chatbots in diverse applications. Chatbots have come a long way from rule-based systems, evolving into advanced AI entities with the ability to comprehend context, emotions, and specialized domains. Breakthrough studies like the examples of GPT-3, DialoGPT, Meena, and DALL-E have showcased significant progress in the capabilities of chatbots.

Constructing a chatbot for a specific domain presents substantial challenges in industry knowledge and methods. It necessitates a profound grasp of context, terminology, and industry trends to deliver precise information. The flexible application of natural language processing and machine/deep learning methods is crucial, utilizing efficient models and algorithms. This paper introduces EduChat, a chatbot system designed to address questions related to universities. EduChat is an artificial intelligence application that combines rule-based methods with novel machine learning techniques to automatically answer common queries related to universities, academic programs, admission procedures, student life, and other related topics. Within EduChat, we have also integrated the ChatGPT system to enhance the coverage of conversation and improve the interaction between humans and machines. To understand the conversations sent to EduChat, we process these sentences in a three-level sequence. First, the text sentences sent to the system are processed by a rule-based system defined by the Definite Clause Grammar (DCG) syntax [5,6]. If the DCG rules fail to recognize the syntactic and semantic structures of the dialogues, an enhanced algorithm built upon the random forest machine learning approach [7] is used to classify these sentences into one of the 25 topic classes. Through experimentation, we establish a classification threshold based on accuracy. Sentences with classification scores below the threshold are processed by ChatGPT within the system's context. Figure 1 depicts a snippet of a conversation from EduChat.

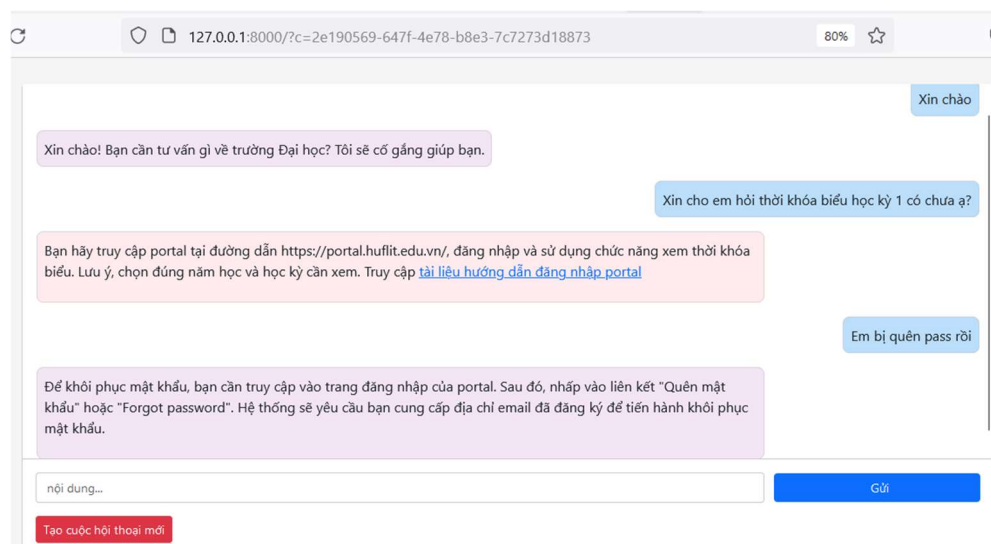


Figure 1. A conversation example with EduChat. (User: Hi EduChat: Hello! I am an AI-based support staff member at the university. How can I assist you? User: Can I ask if the schedule for semester 1 is available yet? EduChat: You can access the portal at the following link: <https://portal.huflit.edu.vn/> (accessed on 1 October 2023). Please log in and use the 'View Schedule' function. Make sure to select the correct academic year and semester you want to view. User: But I forgot my password. EduChat: You can click on 'Forgot Password' on the portal to reset your password.).

The present paper is to explore the following issues:

1. What effective method to build a chatbot to support a specific domain such as issues related to universities?
2. For languages with limited resources such as Vietnamese, does the combination of various methods without relying on deep learning bring effectiveness? The main contributions of this paper are as follows:
 - With only a small dataset in the low-resource Vietnamese language, we have proposed an efficient hybrid method to construct a robust chatbot system for the field of education.
 - Within EduChat, for managing dialogues within the specialized domain, we have established an effective set of rules using DCG and introduced the improved random forest algorithm to enhance text classification accuracy.
 - The model presented in this paper can be applied to various systems, across different contexts and languages.

The remaining sections of this paper are organized as follows: Section 2 provides an overview of related work to contextualize our research. In Section 3, we introduce the proposed model. Sections 4 and 5 detail the components of EduChat. Experimental results will be discussed in Section 6. Finally, this paper concludes in Section 7, summarizing key findings and proposing future directions.

2. Related Work

In this section, we will offer a broad overview of chatbot deployment techniques, including rule-based chatbots and artificial intelligence-driven chatbots (AI).

2.1. Rule-Based Chatbots

Initially, the earliest chatbot implementations relied on rule-based models. These models are typically simpler to design and implement but have certain limitations in terms of functionality. They face challenges when dealing with complex queries. Rule-based chatbots respond to user queries by searching for pattern matches, which makes them prone to providing incorrect answers when encountering sentences that lack recognized patterns.

The advantage of rule-based methods is that, despite the challenges of manually encoding matching rules, they yield high-accuracy results for specific domains and specific scopes. Several noteworthy rule-based chatbot projects include the following: In [8], the authors present a rule-based inquiry chatbot designed exclusively for Asia Pacific University (APU) students, known as the 'APU Admin Bot'. This chatbot is engineered to provide students with a more efficient means of addressing their inquiries, reducing the need for extensive interaction with administrative offices. It functions on a rule-based framework for pattern recognition, whereby specific words, phrases, and actions trigger predefined responses from the chatbot. In [9], a chatbot is described that directs users towards specific actions by employing established rules and decision trees. These scripts and regulations come preconfigured, and any modifications require manual intervention by the organization. One of the initial alternative therapies explored in this project is cognitive behavioral therapy. The authors in [10] introduce an automated method for extracting Artificial Intelligence Markup Language rules from genuine Twitter conversation data.

2.2. Artificial Intelligence-Driven Chatbots

In contrast to rule-based models, AI models rely on machine learning algorithms that enable them to learn from an existing database of human conversations. To achieve this, they require training using machine learning algorithms and a training dataset. With the application of machine learning algorithms, there is no longer a necessity to manually create and code new pattern matching rules. This flexibility liberates chatbots from dependency on domain-specific knowledge. Some exemplary projects applying AI-based methods are as follows. ChatGPT, developed by OpenAI, is an iteration of GPT-3 that has been purpose-built for engaging in conversational interactions, offering articulate and

context-aware responses in natural language. BlenderBot, created by Facebook, is a chatbot that fuses extensive neural models with techniques for multi-turn conversations, with the objective of participating in more substantial and logically connected dialogues [11]. IBM's Project Debater [12] is an artificial intelligence system specialized in debating with humans, proficient at presenting comprehensive and well-structured arguments using natural language. AI-based chatbots have been applied in various fields, ranging from healthcare [13,14], entertainment [15,16], to education [17–20].

3. The Proposed Model

The proposed model aims to improve the understanding and interaction capabilities of the chatbot in complex situations, thereby enhancing the user experience. EduChat operates according to the following mechanism:

1. **Start:** EduChat initiates the conversation with a welcoming message.
2. **User Input:** Users send messages, which can be either questions (queries) or statements.
3. **Understanding and Message Classification:** EduChat comprehends the intent and context of the messages as follows:
 - If the user's input falls within the context of the application:
 - EduChat provides relevant responses by using predefined answers or accessing information from the knowledge base.
 - If the user's input indicates a request beyond the scope of the application:
 - EduChat acknowledges the request and forwards it to ChatGPT to address such queries.
4. **Conclusion:** When user questions have been resolved or when users indicate they have completed their interaction, EduChat concludes the conversation. EduChat bids farewell or provides concluding information.

The following modules are designed to implement the operating mechanism described above (illustrated in Figure 2):

- **User Interface:** A web-based chat window serving as the user interface.
- **Conversation Management:** Manages the flow of the conversation. This ensures that EduChat responds appropriately to user input, adheres to the context, and maintains a smooth conversation.
- **Knowledge Base:** Contains information that EduChat can use to answer user queries. The knowledge base includes frequently asked questions, help guides, and other university-related information.
- **Natural Language Processing (NLP) Module:** This module allows EduChat to understand and process human language. Its role involves parsing the syntax and semantics of text messages.
- **Machine Learning Module:** EduChat is equipped with machine learning and artificial intelligence approaches to comprehend and respond to messages that the NLP module cannot handle.
- **ChatGPT API:** EduChat is integrated with ChatGPT through an API to handle messages evaluated as below the threshold by the machine learning module.

The EduChat model utilizes an NLP module and an improved random forest machine learning module as two main modules to process input data and generate responses. The algorithm of the EduChat model is described as follows (Algorithm 1).

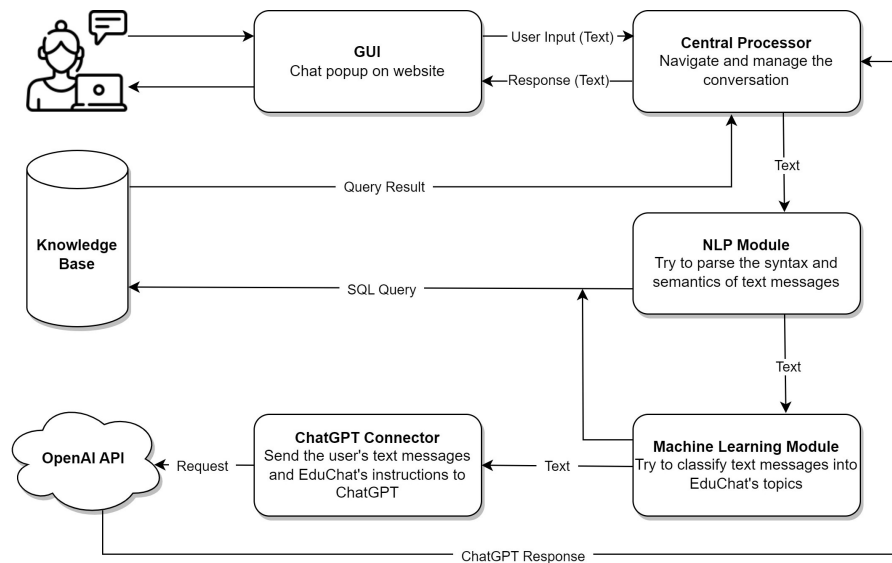


Figure 2. EduChat system architecture.

Algorithm 1. EduChat Algorithm

function EduChat(conversation)

Input: conversation

Output: response content for the user

Variables:

user_input: the newest user's message

prolog: an instance of the Prolog class. DCG syntax and semantics rules was loaded into this variable.

model: an instance of the improved random forest (IRF) model.

db: database object.

gpt: an instance of ChatGPTConnector, which helps connect to ChatGPT via Open AI's APIs.

```

1. user_input = encode(user_input) # user_input in Vietnamese will be preprocessed, standardized
   and encoded.
2. task = prolog.query(user_input) # try to parse the syntax and semantics of user input
3. if task is None then
4.     predicted_task, prob = model.predict_prob(user_input) # The IRF model will predict
   and give the probability of each class. If the maximum probability value is greater than the system's
   configuration threshold, the system will accept this classification value.
5.     if prob ≥ threshold then
6.         task = predicted_task
7. if task is not None then
8.     response = db.query(task)
9. else
10.    response = gpt.post(conversation, user_input)
11. return response
end function

```

4. Natural Language Processing (NLP)

NLP plays a crucial role in chatbot development. It enables the chatbot to understand and interact with human language naturally, thereby improving its ability to respond and understand user intent. Thanks to NLP, a chatbot can handle a wide range of questions and requests, providing users with a smooth and efficient interaction experience. We conducted a survey and collected 625 text messages from exchanges with students involving 06 academic advisors and 02 student affairs staff members from HUFLIT University and

Tien Giang University. These messages were categorized into 25 topics, and some typical message types are presented in Table 1.

Table 1. Some query types and semantic presentations classified by topic.

ID	Department	Topic	Example	Semantic Presentation
1	Công tác sinh viên(The Student Affairs Office)	Xác nhận sinh viên (Student confirmation letter)	Chào phòng Công tác sinh viên, em muốn làm giấy xác nhận sinh viên. (English: Hello Student Affairs Office, I would like to request a student confirmation letter.)	query (dept (ctsv), task (xacnhan))
		Kết quả rèn luyện(The training results)	Phòng Công tác sinh viên hướng dẫn em đánh giá kết quả rèn luyện với ạ? (English: Student Affairs Office, could you please guide me on how to evaluate my training results?)	query (dept (ctsv), task (renluyen))
		Thẻ sinh viên (Student ID card)	Phòng Công tác sinh viên cho em xin được cấp lại thẻ sinh viên ạ? (English: Student Affairs Office, could you please issue me a replacement student ID card?)	query (dept (ctsv), task (thesinhvien))
		Mùa hè xanh ("Green Summer Campaign")	Phòng Công tác sinh viên cho em thông tin về chiến dịch mùa hè xanh năm nay ạ? (English: Student Affairs Office, could you please provide me with information about this year's "Green Summer Campaign"?)	query (dept (ctsv), task (muahexanh))
		Hiến máu nhân đạo (Voluntary blood donation)	Phòng Công tác sinh viên cho em thông tin về hiến máu nhân đạo ạ? (English: Student Affairs Office, could you please provide me with information about voluntary blood donation?)	query (dept (ctsv), task (hienmau))
		Công tác xã hội (Social work)	Kính gửi Phòng Công tác sinh viên, em muốn tham gia công tác xã hội ạ. (English: Dear Student Affairs Office, I would like to participate in social work.)	query (dept (ctsv), task (hienmau))
2	Đào tạo (The Department of Training)	Đăng ký môn học (Register for courses)	Phòng Đào tạo cho em hỏi cách đăng ký môn học ạ. (English: The Department of Training, could you please tell me how to register for courses?)	query (dept (daotao), task (dangkymonhoc))
		Thời khóa biểu (Class schedule)	Phòng Đào tạo cho em hỏi xem thời khóa biểu ở đâu ạ. (English: The Department of Education and Training, where can I check the class schedule, please?)	query (dept (daotao), task (thoikhoabieu))
		Lịch thi (Exam schedule)	Phòng Đào tạo cho em hỏi xem lịch thi ở đâu ạ. (English: The Department of Training, where can I find the exam schedule, please?)	query (dept (daotao), task (lichthi))
		Bảng điểm (Grade report)	Phòng Đào tạo cấp cho em bảng điểm ạ. (English: The Department of Training, could you please provide me with a grade report, please?)	query (dept (daotao), task (bangdiem))
		Bảo lưu (A leave of absence)	Em liên hệ với phòng Đào tạo để bảo lưu học kỳ này. (English: I contact the Department of Training to request a leave of absence for this semester.)	query (dept (daotao), task (baoluu))
		Chuyển ngành (A change in major)	Em liên hệ với phòng Đào tạo để muốn chuyển ngành. (English: I will contact the Department of Training to request a change in major.)	query (dept (daotao), task (chuyennghanh))
		Tốt nghiệp (Graduation)	Xin phòng Đào tạo cho em giấy chứng nhận tốt nghiệp tạm thời ạ. (English: Please provide me with a temporary graduation certificate, thank you.)	query (dept (daotao), task (totnghiep))

Table 1. Cont.

ID	Department	Topic	Example	Semantic Presentation
3	Tài vụ(The Finance Office)	Đóng học phí (Tuition fee)	Phòng Tài vụ cho em hỏi học phí học kỳ này là bao nhiêu tiền ạ. (English: The Finance Office, may I ask how much the tuition fee is for this semester, please?)	query (dept (taivu), task (tienhocphi))
		Học phí trễ, gia hạn học phí (tuition fee payments)	Phòng Tài vụ còn nhận đóng học phí không ạ. (English: The Finance Office, are you still accepting tuition fee payments?)	query (dept (taivu), task (hocphitre))
4	Khoa chuyên môn (Faculty)	Đăng ký phúc khảo (Reevaluation)	Xin cho em thủ tục làm đơn phúc khảo môn Mạng máy tính. (English: Please provide me with the procedures for requesting a reevaluation of the Computer Networking subject.)	query (dept (khoa), task (phuckhao))
		Xin giấy giới thiệu (letter of introduction)	Em muốn làm giấy giới thiệu của trường để đi thực tập ạ. (English: I would like to request a letter of introduction from the school for my internship, please.)	query (dept (khoa), task (giaygioithieu))
5	Tuyển sinh (Admission)	Chỉ tiêu tuyển sinh (admission quota)	Xin cho em biết chỉ tiêu tuyển sinh năm nay là bao nhiêu? (English: Could you please inform me of this year's admission quota?)	query (dept (tuyensinh), task (chitieu))
		Ngành học (Major)	Trường tuyển sinh những ngành nào? (English: What majors is the school admitting students into?)	query (dept (tuyensinh), task (nganh))
		Điểm chuẩn (Cutoff score)	Điểm chuẩn của trường là bao nhiêu? (English: What is the school's admission cutoff score?)	query (dept (tuyensinh), task (diemchuan))
		Phương thức xét tuyển (Admission methods)	Trường có những phương thức xét tuyển nào? (English: What admission methods does the school offer?)	query (dept (tuyensinh), task (phuongthucxettuyen))
		Khối thi (Subject groups)	Trường có những khối thi nào? (English: What subject groups does the school have for the entrance exam?)	query (dept (tuyensinh), task (khoi))
6	Chung (Others)	Xin thư giới thiệu (letter of recommendation)	Em muốn xin thư giới thiệu thì liên hệ đâu ạ? (English: If you want to request a letter of recommendation, where should you contact, please?)	query (contact (thugioithieu))
		Cấp lại thẻ sinh viên (Student ID card)	Em muốn cấp lại thẻ sinh viên thì liên hệ phòng nào ạ? (English: If I want to request a replacement student ID card, which department should I contact?)	query (contact (thesinhvien))
		Miễn giảm học phí (Tuition fee exemption)	Em cần làm đơn miễn giảm học phí thì liên hệ với ai? (English: To apply for a tuition fee exemption or reduction, who should I contact?)	query (contact (mienhocphi))

The scope of the EduChat application includes clear rules and logic, and situations that can be predicted. The data within EduChat is also limited and straightforward. This helps us choose a method based on a set of rules and conditions to handle every situation.

To represent the semantics of the text messages, we use DCG [5,6], also known as “Prolog-style grammars”. DCG is a method in NLP and computational linguistics that employs logic-based grammar. DCG is built upon the grammar of logical propositions. Each proposition typically consists of a head and a body, where the head contains a statement, and the body contains the necessary conditions for that statement to be satisfied. DCG is a powerful tool in the field of NLP and is suitable for building chatbot systems.

Example 1: “Phòng Công tác sinh viên cho em hỏi thủ tục làm giấy xác nhận sinh viên như thế nào?” (English: The Student Affairs Office, I would like to ask how to go about obtaining a student certificate?)

The DCG syntax and semantics rules are defined as follows:

```
query (query (Dept, Task)) --> n_dept, n_dept (Dept), n_phrase, v_how, n_task (Task), w_how.
n_dept --> [phòng].
n_dept (dept(Dept)) --> n_depts (Dept).
n_phrase --> [cho, em, hỏi].
n_depts (ctsv) --> [công, tác, sinh, viên].
v_how --> [thủ, tục, làm].
n_task (task(Task)) --> n_tasks(Task).
n_tasks (xacnhan) --> [giấy, xác, nhận, sinh, viên].
w_how --> [như, thế, nào].
```

We obtain the syntax and semantics rules that define the semantic structure of query example 1 as follows: query (dept (ctsv), task (xacnhan)).

From these semantic structures, we proceed to convert them into corresponding SQL queries for database/knowledge base retrieval. The results returned from the database/knowledge base are processed to provide responses to the user.

5. Intent Classifier Based on Improved Random Forest Algorithm (Machine Learning Module)

We propose an intent classifier based on improved random forest to handle cases where the NLP module described in Section 4 cannot resolve incoming messages. In such cases, the messages are automatically classified into one of the 25 groups corresponding to the 25 semantic structures presented in Table 1. To ensure sufficient data for applying machine learning methods, we perform data augmentation as presented by Section 5.1.

5.1. Data Augmentation

Data augmentation is a technique introduced to enhance the accuracy of training deep learning models by expanding the dataset necessary for the training process. For instance, the work of Wei et al. [21] introduced data augmentation techniques that utilize internal data rather than relying on language models and external sources. These techniques encompassed synonym replacement, which substitutes a specific word in a sentence with a synonym; random insertion, which adds a random word; random swap, which exchanges the positions of two random words in a sentence; and random deletion, which removes a random word.

We conducted text augmentation based on the idea from Wei et al. [21] and obtained a dataset comprising 6250 text sentence samples categorized into 25 classes. The data preprocessing steps included tokenization, stop word removal, and feature extraction using TF-IDF.

5.2. The Improved Random Forest (IRF) Algorithm

Random forest is a powerful machine learning algorithm introduced by Leo Breiman and Adele Cutler in 2001 [7]. It quickly became an important tool in building predictive models and analyzing data. Random forest works by creating an ensemble of multiple random decision trees. Each decision tree is trained on a bootstrap sample, which is a subset of the original data. This helps reduce the risk of overfitting, a common issue in machine learning. When making predictions, random forest aggregates the results from all the decision trees to provide the final prediction. Random forest has several advantages, including its ability to handle noisy and large datasets, its applicability to classification and prediction tasks, and its capability to estimate the importance of input variables. The equation for defining the mean square error (MSE) of a random forest is as follows.

$$\text{MSE} = \frac{1}{N} \sum_{k=0}^n \binom{n}{k} (F_i - F_i) \cdot b^2 \quad (1)$$

In Equation (1), N represents the total count of unique data points, where F_i represents the predicted outcome provided by model Y_i , and i corresponds to the specific data point in question.

In the system, the IRF algorithm consists of 02 main phases: training and testing.

Training phase. During the training phase, the standard random forest algorithm is enhanced by introducing a cost function that relies on the misclassification rate of the samples.

The misclassification rate of the samples is calculated by assessing the false negative (FN) and false positive (FP) values in the confusion matrix for each decision tree during the training process.

$$\text{cost}(T) = (\text{FN}(T) + \text{FP}(T)) / |\text{size}(T)| \quad (2)$$

where $\text{FN}(T)$ represents the count of false negatives in a tree T , $\text{FP}(T)$ signifies the count of false positives in tree T , and $|\text{size}(T)|$ denotes the number of samples in the training set of tree T .

Subsequently, an error-based technique is devised to allocate weights to each tree. This error is computed in 02 components: the overall training error and the positive training error.

The overall training error is computed by considering the total misclassification of samples within a tree. The overall positive error is computed based on the minority class samples in a tree. In detail, the error calculation is defined as:

$$\text{err}_{ov}(x_i) = \begin{cases} 1 & \text{if } y(x_i) \neq \text{class}(x_i) \\ 0 & \text{if } y(x_i) = \text{class}(x_i) \end{cases} \quad (3)$$

$$\text{err}_{pos}(x_m) = \begin{cases} 1 & \text{if } y(x_m) \neq \text{positive} \\ 0 & \text{if } y(x_m) = \text{positive} \end{cases} \quad (4)$$

where x_i is a sample in a tree, $y(x_i)$ is the class of x_i assigned by the tree, and x_m is a minority class sample in the tree.

If more than one tree exhibits the same overall training error, the tree with the lowest overall positive error is assigned a higher weight.

The cost function reflects the predictive capability of the tree, while the weight formulation indicates the confidence level associated with a tree. The optimal tree is determined by selecting the one with both the highest predictive ability and the highest confidence value.

The complete training process is outlined in the Algorithm 2.

Algorithm 2. IRF_1

Input: S is dataset, S_{min} is minority class set, S_{maj} is majority class set, n : number of trees.

Output: Weighted trees' set, $\{\text{weight}(T): T \text{ is a tree}\}$

1. Split the set S into *train* and *test*.

2. By sampling with replacement, divide *train* into n training sets, termed as T .

3. For $i := 1$ to n do

(a) Split T into T_{maj} and T_{min} : $T_{maj} \subseteq S_{maj}$, $T_{min} \subseteq S_{min}$.

(b) Train T with decision tree learning algorithm.

(c) Compute the overall error using Equations (2) and (3) as:

$$\text{err}_{ov}(T) = \sum_{m=1}^N \text{err}_{ov}(x_i) * \text{cost}(T)$$

N = total number of samples in the tree T .

(d) Compute the positive training error, using Equations (2) and (4) as

$$\text{err}_{pos}(T) = \sum_{m=1}^p \text{err}_{pos}(x_m) * \text{cost}(T)$$

p is total number of minority instances in the tree T .

(e) Assign the tree T_i with $\text{weight}(T_i) = 1/\text{err}_{ov}(T_i)$

Tree with lower error will obtain higher weight.

(f) In case that the overall training errors of the two or more trees are the same, the weights are assigned using the formula:

$$\text{weight}(T_i) = 1/\text{err}_{pos}(T_i)$$

Testing phase. During the testing phase, the result for each sample in the test set is determined by assessing it against each of the trained decision trees and then aggregating

the outcomes obtained from all the trees. In a standard random forest, the final result or label for a test sample is established by considering the majority vote from all the outcomes generated by the individual trees. However, in the case of IRF, the final outcome is decided based on the weight assigned to a specific selected tree. For a given test pattern, x' the final outcome, $F(x')$ is given as:

$$F(x') = f(x') \text{ from } T_i \mid \text{weight}(T_i) = \text{highest} \quad (5)$$

In other words, the result obtained from the tree with the highest weight for a given test sample is taken as the ultimate outcome.

The entire testing process is depicted in Algorithm 3.

Algorithm 3. IRF_2

Input: *test* is the set of test samples, $W = \{\text{weight}(T)\}$ is the set of weights of all trees T .

Output: Classified test set

1. Classify *test* with all the trees
 2. For each sample x' in *test*
 - (a) for each tree T_i in the forest obtains the outcome label as $y(x') = f(x')$ from T_i
 - (b) determine the final outcome, $F(x')$ by using Equation (5).
-

One distinguishing property of IRF is that the training process of each tree in this random forest model can be performed independently.

6. Experimentation and Evaluation

We sequentially conducted tests for the NLP module and the classifier module using IRF method. Following that, we conducted a survey to evaluate user perceptions of the system.

6.1. The NLP Module

With the aim of testing the capability to handle clear rules and logic, as well as predictable situations of the NLP module in the proposed model, we conducted tests on a total of 50 text messages sent to the NLP module. The results indicate that this module operates accurately for all 50 test messages. These messages belong to a series of predefined syntax and semantics structures for the system. The ability to accurately process all these standard sentences demonstrates the stability and accuracy of the NLP module. Some test results are provided by Table 2.

Table 2. Some selected test results on the NLP module.

Text Messages	Results
Em muốn làm giấy xác nhận sinh viên. (English: Hello, I would like to request a student confirmation letter.)	query (dept (ctsv), task (xacnhan))
Xin cho em hỏi mình xem thời khóa biểu ở đâu ạ? (English: Could you please tell me where can I check the class schedule?)	query (dept (daotao), task (thoikhoabieu))
Chỉ tiêu tuyển sinh năm nay là bao nhiêu sinh viên? (English: How many students is the admission target for this year?)	query (dept (tuyensinh), task (chitieu))
Xin chào Phòng Đào tạo, em muốn được cấp bảng điểm. (English: Hello Academic Affairs Office, I would like to request my exam score transcript.)	query (dept (daotao), task (bangdiem))

Coverage Scope: If there are cases where sentences do not fit into the predefined syntax structures, the system will return inaccurate parsing results. This indicates that the DCG syntax rules and built-in dictionary are not broad enough to encompass all cases. If supplemented with a larger lexicon and more comprehensive DCG rules, the coverage scope of the proposed model will significantly improve.

6.2. Intent Classifier

The Intent Classifier is a crucial module in the proposed model. Sentences outside the coverage scope of the NLP module will be classified by the Intent Classifier. We split the dataset in an 8:2 ratio for training and testing. The training and testing processes were conducted using machine learning methods including improved random forest (IRF), K Nearest Neighbors, Support Vector Machines, decision tree, Neural Net, as well as deep learning methods such as Long Short-Term Memory Networks (LSTM) [22], Bidirectional Long Short-Term Memory Networks (BiLSTM) [23]. Table 3 and Figure 3 present the results of these experiments. The IRF method proved to be superior to other methods with an accuracy of up to 97.69%.

Table 3. Experimental results with 08 machine learning and deep learning methods.

ID	Methods	Accuracy
1	K Nearest Neighbors (KNN)	0.8692
2	Support vector machines (SVM)	0.9692
3	Decision Tree (DT)	0.9308
4	Improved Random Forest (IRF)	0.9769
5	Neural Net (ANN)	0.9538
6	Naive Bayes (NB)	0.9538
7	LSTM [22]	0.9688
8	BiLSTM [23]	0.9698

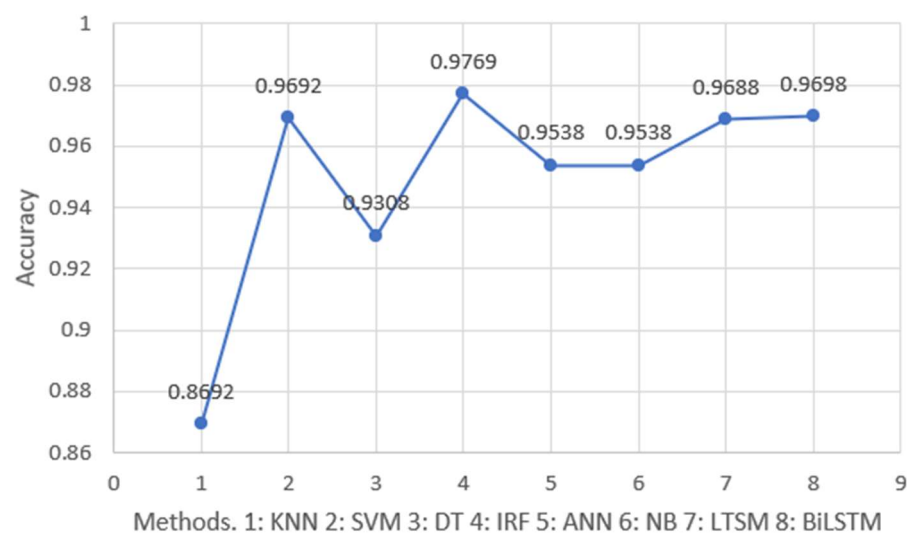


Figure 3. Classification results using machine learning and deep learning methods on the EduChat system.

6.3. System Testing and Discussion

We also conducted surveys with EduChat users using Technology Acceptance Model method [24], asking them, “Is the system easy to use?” with four rating levels (Very convenient, Quite convenient, Somewhat convenient, Not convenient), and “Does the chatbot provide useful information and solve a significant portion of your problems? With four rating levels (Very useful, Quite useful, Somewhat useful, Not useful)”. These surveys were conducted from 23 September 2023, to 3 October 2023 and 28 October 2023 to 30 October 2023, for 40 first-year students at the beginning of the first semester at HUFLIT University. The evaluation level “convenience” ranges from fairly to very convenient, accounting for 85%, corresponding to 34 out of 40 survey participants, and the evaluation

level “useful” ranges from fairly to very useful, accounting for 90%. The results are shown in Table 4.

Table 4. Survey of the usability level of the system.

Very Convenient	Quite Convenient	Somewhat Convenient	Not Convenient
40%	45%	10%	5%
Very useful	Quite useful	Somewhat useful	Not useful
42.5%	47.5%	5%	5%

The participation of the system’s modules was also considered in the statistics. Out of 128 messages sent to the system, 11 were processed by the NLP module, 98 messages were classified by the machine learning module, and ChatGPT participated in processing 19 messages. Furthermore, we also conducted an analysis of 19 messages that ChatGPT processed. These were general messages about psychological issues, study methods, etc., and ChatGPT provided good advice on these issues. The integration of ChatGPT has made the system more flexible and diverse.

From the results above, it can be seen that EduChat is a robust system based on the hybrid combination of both rule-based and machine learning methods. The rule-based approach is highly reliable, easy to verify and interpret, capable of handling limited data, and works well with incomplete data. While machine learning has the capability to learn and improve autonomously from data, achieving high performance in processing large and complex datasets, it can uncover complex patterns and adapt to diverse data without requiring expert knowledge beforehand. Furthermore, EduChat integrates ChatGPT’s API, leveraging all the capabilities of ChatGPT to handle every conversation input to the system effectively.

Currently, predictions for the future of chatbots show a strong trend of continued development. Some potential trends for the future of chatbots may include the integration of artificial intelligence (AI) and machine learning, multi-channel interaction, the development of versatile chatbots, stronger security integration, and improved user experience integration.

7. Conclusions

Despite the current advancements in the fields of NLP, the capabilities of chatbots still have limitations, with a notable example being that these chatbots cannot provide up-to-date information or answers related to events or news. To some extent, in order to address that situation, this paper has presented the architecture and approach used to develop the EduChat system. In EduChat, the NLP and machine learning modules play crucial roles in message processing. To our knowledge, this is the first system in Vietnam that integrates NLP mechanisms, effective machine learning methods, and ChatGPT, creating intelligence and flexibility in the system. The encouraging results of this paper have answered the two research questions posed. This research also opens up new opportunities for the development and expansion of automated conversational systems capable of understanding and communicating in Vietnamese with users. In the future, we plan to train on large datasets for the machine learning module and define additional rule sets for the natural language processing component to enhance the system’s flexibility and accuracy.

Author Contributions: Writing—original draft, H.D.; writing—review and editing, T.K.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting this study’s findings are available from the corresponding author upon reasonable request. The data are not publicly available due to privacy.

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

References

1. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20), Vancouver, BC, Canada, 6–12 December 2020.
2. Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; Dolan, B. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics (ACL), Online, 5–10 July 2020; pp. 270–278. [CrossRef]
3. Adiwardana, D.; Luong, M.-T.; So, D.R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. Towards a Human-Like Open-Domain Chatbot. *arXiv* **2020**, arXiv:2001.09977v3.
4. Reddy, M.D.M.; Basha, M.S.M.; Hari, M.M.C.; Penchalaiah, M.N. Dall-e: Creating images from text. *UGC Care Group I J.* **2021**, *8*, 71–75.
5. Pereira, F.C.N.; Shieber, S.N. *Prolog and Natural-Language Analysis*; Microtome Publishing: Brookline, MA, USA, 2005.
6. Tran, T.K.; Nguyen, D.T. Semantic processing mechanism for listening and comprehension in VNCalendar system. *Int. J. Nat. Lang. Comput. (IJNLC)* **2013**, *2*, 1–15. [CrossRef]
7. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
8. Singh, J.; Joesph, M.H.; Jabbar, K.B.A. Rule-based chatbot for student enquiries. *J. Phys. Conf. Ser.* **2019**, *1228*, 012060. [CrossRef]
9. Rath, S.; Pattanayak, A.; Tripathy, S.; Priyadarshini, S.B.B.; Tripathy, A.; Tanvi, S. Prediction of a Novel Rule-Based Chatbot Approach (RCA) using Natural Language Processing Techniques. *Int. J. Intell. Syst. Appl. Eng.* **2023**, *11*, 318–325. Available online: <https://www.ijisae.org/index.php/IJISAE/article/view/3172> (accessed on 17 September 2023).
10. Yamaguchi, H.; Mozgovoy, M.; Danielewicz-Betz, A. A Chatbot Based On AIML Rules Extracted From Twitter Dialogues. In Proceedings of the Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems, PTI, Poznań, Poland, 9–12 September 2018; pp. 37–42. [CrossRef]
11. Shuster, K.; Xu, J.; Komeili, M.; Ju, D.; Smith, E.M.; Roller, S.; Ung, M.; Chen, M.; Arora, K.; Lane, J.; et al. BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage. *arXiv* **2022**, arXiv:2208.03188v3.
12. Slonim, N.; Bilu, Y.; Alzate, C.; Bar-Haim, R.; Bogin, B.; Bonin, F.; Choshen, L.; Cohen-Karlik, E.; Dankin, L.; Edelstein, L.; et al. An autonomous debating system. *Nature* **2021**, *591*, 379–384. [CrossRef] [PubMed]
13. Solanki, R.K.; Rajawat, A.S.; Gadekar, A.R.; Patil, M.E. *Building a Conversational Chatbot Using Machine Learning: Towards a More Intelligent Healthcare Application*; IGI Global: Hershey, PA, USA, 2023; pp. 285–309. Available online: <https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-6684-7164-7.ch013> (accessed on 1 October 2023).
14. Pandey, S.; Sharma, S.; Wazir, S. Mental healthcare chatbot based on natural language processing and deep learning approaches: Ted the therapist. *Int. J. Inf. Technol.* **2022**, *14*, 3757–3766. [CrossRef]
15. Garcia-Mendez, S.; De Arriba-Perez, F.; Gonzalez-Castano, F.J.; Regueiro-Janeiro, J.A.; Gil-Castineira, F. Entertainment Chatbot for the Digital Inclusion of Elderly People without Abstraction Capabilities. *IEEE Access* **2021**, *9*, 75878–75891. [CrossRef]
16. De Arriba-Pérez, F.; García-Méndez, S.; González-Castaño, F.J.; Costa-Montenegro, E. Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with Natural Language Processing capabilities. *J. Ambient Intell. Humaniz. Comput.* **2022**, *1*, 1–16. [CrossRef] [PubMed]
17. Mageira, K.; Pittou, D.; Papasalouros, A.; Kotis, K.; Zangogianni, P.; Daradoumis, A. Educational AI Chatbots for Content and Language Integrated Learning. *Appl. Sci.* **2022**, *12*, 3239. [CrossRef]
18. Lee, D.; Yeo, S. Developing an AI-based chatbot for practicing responsive teaching in mathematics. *Comput. Educ.* **2022**, *191*, 104646. [CrossRef]
19. Görtz, M.; Baumgärtner, K.; Schmid, T.; Muschko, M.; Woessner, P.; Gerlach, A.; Byczkowski, M.; Sülthmann, H.; Duensing, S.; Hohenfellner, M. An artificial intelligence-based chatbot for prostate cancer education: Design and patient evaluation study. *Digit. Health* **2023**, *9*, 20552076231173304. [CrossRef] [PubMed]
20. Chen, Y.; Jensen, S.; Albert, L.J.; Gupta, S.; Lee, T. Artificial Intelligence (AI) Student Assistants in the Classroom: Designing Chatbots to Support Student Success. *Inf. Syst. Front.* **2023**, *25*, 161–182. [CrossRef]
21. Wei, J.; Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 6382–6388. [CrossRef]
22. Khataei, M.H.; Gharehchopogh, F.S.; Majidzadeh, K.; Sangar, A.B. A New Hybrid Based on Long Short-Term Memory Network with Spotted Hyena Optimization Algorithm for Multi-Label Text Classification. *Mathematics* **2022**, *10*, 488. [CrossRef]
23. Huan, H.; Guo, Z.; Cai, T.; He, Z. A text classification method based on a convolutional and bidirectional long short-term memory model. *Connect. Sci.* **2023**, *34*, 2108–2124. [CrossRef]
24. Davis, F. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.