International Journal of
Emerging Multidisciplinary Research

ISSN : 2546-1583
Volume 4, Number 2, June 2020
https://doi.org/10.22662/IJEMR.2020.4.2.013

# Development of an AI Chatbot to Support Admissions and Career Guidance for Universities

**Le Hoanh Su[1], Truong Dang-Huy[2], Tran Thi-Yen-Linh[2], Nguyen Thi-Duyen-Ngoc[2], Ly Bao-Tuyen[2], Nguyen Ha-Phuong-Truc[3]**

[1]Faculty of Information Systems, University of Economics and Laws, VNU-HCM, Ho Chi Minh City, Vietnam
[2]Department of Ecommerce, Faculty of Information Systems, University of Economics and Laws, VNU-HCM, Ho Chi Minh City, Vietnam
[3]Department of Management Information Systems, Faculty of Information Systems, University of Economics and Laws, VNU-HCM, Ho Chi Minh City, Vietnam

***Abstract***

Background/Objectives: The thesis "Building AI Chatbot to support admissions and career guidance for universities" has been explored, analyzed and understood the difficulties, shortcomings and problems arising in career counseling, enrollment support. Methods/Statistical analysis: Building a chatbot is a perfect solution. A chatbot can be available 24/7 for 365 days. Chatbot also receives and processes requests of students / parents automatically and quickly and uniformly in replying content, especially optimized in repetitive scenarios. Findings: Thus we decided to do research to understand this situation. Then creating a dataset supports vocational guidance and advising education enrollment activities. We also design and integrate chatbot into the school system to support the admissions counseling process. Improvements/Applications: Besides, the thesis has successfully built a structured dataset about enrollment orientation and apply of natural language processing, machine learning to build identification models. The thesis will contribute to overcome and improve the performance of university admissions consulting.

***Keyword***

Admissions, chatbot, AI, data, machine learning, information systems.

**Corresponding author : Le Hoanh Su**
sulh@uel.edu.vn

• Manuscript received April 25, 2020.
• Revised May 20, 2020 ; Accepted June 20, 2020.
• Date of publication June 30, 2020.

# I. INTRODUCTION

The development of artificial intelligence is going beyond what people can imagine, chatbot is increasingly asserting its important role in many areas of life. Chatbot is being researched and developed at a rapid pace. Weiyu Wang et al (2018) [1] said that chatbot has been used to develop and advance numerous fields and industries, including finance, healthcare, education, transportation, and more.

At present, the application of chatbot in career guidance and admissions consulting is increasingly attracting the attention of universities and colleges. The construction and completion of chatbot tools based on information technology and enrollment data, career counseling data contributes to the effective support for enrollment counseling at universities and colleges throughout the country. Chatbot can solve difficult problems in online enrollment counseling.

For example, it takes a lot of time and manpower to answer simple and similar questions, meeting processing speed limits and errors caused by typing, spelling, expression, distraction when being affected by the surroundings,... At the same time, many universities and colleges offer 4.0 majors in digital technology and artificial intelligence, which is the industry that captures the needs and catches the market trend during the 4th industrial revolution.

Believe that artificial intelligence will be a hot industry in the future. However, the understanding of scholars and parents about this industry is not deep, there are still many questions and difficulties that require career support and advice. For the career counseling team, they are necessary to cultivate knowledge of the industry, they need to be trained in an easy-to-understand and appropriate interpretation method to specifically address the students and parents. This will take a lot of time and effort to train an entire team of consultants, and the risks of lack of experience will also appear a lot.

Building a chatbot is a perfect solution. A chatbot can be available 24/7 for 365 days. Chatbot also receives and processes requests of students / parents automatically and quickly and uniformly in replying content, especially optimized in repetitive scenarios. This will help to save a part of manpower, improve the quality of advice, limit the errors that arise and increase experience when interacting with the university, so it contributes to improving the reputation, quality and image of the University for candidates, parents.

The establishment of the enrollment counseling data set, answering questions, designing, programming and integrating chatbot into the school information system to support the admissions counseling process and solve problems that are encountered in the work of admissions counseling at universities and colleges. Then offering optimal solutions for admissions department to support for candidates, parents to access and understand the enrollment information in the best way.

# II. THEORETICAL FRAMEWORK

## *A. Literature Review*

We refer to the chatbot model of Massimiliano et al (2018) [2], they built a Virtual Assistant to Help Students in Their University Life, it been built to assist students with issues related to university life and research how Chatbot's personality affects user experience, combined with surveys to understand user needs and behavior.

In addition, Mohammed Ismail and Abejide Ade-Ibojola (2019) [3] also recognize similar problems so they decide to build a Chatbot as an assistant lecturer to support for students who have difficulty in subjects related to programming, algorithms and logic. The authors also analyzed Chatbot's operational model, the method of interacting, sending requests and processing requests on the server, methods of data normalization and feedback.

Along with the research direction on how to build Chatbot, in the article of Patrick Bii et al (2013) [4] - "Investigation of student's attitude towards use of chatbot technology in Instruction: the case of Knowie in a selected high school" published in Educational Research, they built Knowie Chatbot on Ubuntu platform to study students' attitudes during learning involved in experience with chatbot, using Python language, JDK software and PyAIML library to build Knowie Chatbot.

Lai Thi Hai Yen et al (2018) [5] have clearly presented on school’s procedures and regulations which are regularly concerned by students and how to build the optimal database.

We also refer to the idea of building an chatbot as a personal assistant that can serve as human’s resume by Gayatri Nair et al (2018) [6].

## *B. Application of theories*

**Analyzing and mining data from HTML page**

In this research, we used methods of analyzing and mining content on websites and social networks. We used the programming language to process HTML strings and pages to get data. By identifying the partial layout and tag structure, we can use some basic string splitting methods to get the content from the webpage.

During the construction of the dataset for ChatBot, we retrieved data by analyzing HTML pages. Rely on the data analyzed from the HTML website to proceed to complete and build the database for ChatBot.

**Apply Text Mining and Natural Language Processing to build dataset**

Text mining includes basic steps such as: preprocess, model learning, prediction, analyzing and presenting results (Dr. S. Vijayarani et al 2015 [7]). In this research, we used Use preprocessing tools and algorithms from the text extraction steps. We also applied Text Mining and Natural Language Processing that we refer the steps of Hoanh-Su Le et al (2017) [8] to build a dataset that is capable of identifying and processing languages from the user to ChatBot and vice versa.

**Using TF-IDF Algorithm**

In the process of processing data sets and building Chatbot's identification model, we refer article of Bijoyan Das et al (2018) [9] used the TF-IDF weighting method to determine how importance of each word in the text and the value of the word for the context and meaning of the sentence.

Words which have high TF-IDF value appear a lot in this text, but less in other texts. This helps to filter out common words and retain high value words (keywords of that text).

**Word vectorization**

It's a special technique in natural language processing, which maps words or phrases from the vocabulary to a corresponding vector of real numbers (Karol Grzegorczyk. 2018 [10]). In this research, we vectorized each sentence, not the whole paragraph. Because if we vectorized in a long paragraph, it may have too many dimensions, leading to inaccuracy, difficult to handle. The vectorization method has 2 ways:

• Using one-hot
• Showing dispersion

**Using SVM model in Classification problem**

Applied model of Steven Kester Yuwono et al (2018) [11] to separate the topics to support the addition of new datasets, we used SVM model to calculate the distance from inputs to boundaries, thereby measuring the correlation between the sentences in each topic and determine which topics input belongs to.

**We built ChatBot based on the two models:**

*Machine learning model for classifying user intent*

First, we conducted data processing, perform "cleaning" data such as removing redundant information, standardizing data such as correcting misspelled words, standardize the abbreviations, etc. Using the model to classify intent, and use intent to classify a new conversation.

*Model-Based Matching*

Refer to the model of Nicole Radziwill et al (2017) [12], in model-based matching, we have a built-in database set, where each intent has at least one corresponding question. With a given conversation, we will apply this method to compare each question in the dataset. The answers to questions that are closest to the input will be given.

**Object-oriented programming method and Chatbot model building**

We conducted building based on object-oriented programming method (OOP) with Python (Max Leuthäuser 2010 [13]). The objects in the Chatbot model are a combination of code, properties, execution functions, and data. Each object has a unique name and all references to that object are processed by its name. Therefore, each object has the ability to receive messages, process data (inside of it), and send or reply to other objects or the environment in OOP.

## III. RESEARCH METHODOLOGY

### *A. Data Collection*

The Chatbot's main task is to answer the questions related to admissions counseling for students and parents. Therefore, the information given should be accurate, official and reliable. To make data sets cover all of possible situations, we have to select two different data sources below:

• The sets of questions are asked by students and parents from the groups on social network of the University of Economics and Law.

• The sets of questions and database are supported by CCA department.

For the data from social media, we refer to data collection model of Hoanh Su Le et al (2015) [14] to download the content HTML file from the groups, then we use Python and the BeautifulSoup, urllib libraries to extract the content and remove the HTML components. Next, we conduct reading comprehension and omitting the redundant content, characters, then storing data.

In addition, we have two data files which are collected during the process of admissions consulting

from including the list of questions and interrogators' information and the list of keywords at different levels. These two lists are enclosed with some topics and content that are often interested in.

### *B. Building data sets*

After extracting a set of questions, we synthesize and systemize it a list of the most common topics and content. We develop a comprehensive list of question sets that contain the corresponding topics and content for each topic. Next, we conduct drafting and create questions with similar meanings.

In order to proceed to compile equivalent questions according to the original question collected, we identify principles to create equivalent questions based on the adjustment of the structure and wording of the sentence.

• Change words in sentences with synonyms
• Change the sentences to active or passive form
• Swap the components of the questions
• Change the question form.

After building a set of questions, we determine the database structure used for the model:

• Questions: A collection of questions, problems, difficulties and complaints entered by the students or parents.
• Answers: A set of responses that correspond to the questions the chatbot receives.
• Additional information for the answers: Information that varies from time to time such as "benchmark", "tuition", "admission date", ... supplementary information pack will act as turn in the answer.
• Other multimedia files: These are images, audio clips, PDF files, etc. that can be used to attach to answers in the consultation and troubleshooting process for users.

After completing the questionnaire set, the structure of the data sheets and the answer set, the research team should carry out the task of cross-checking and editing the data set to avoid unnecessary errors. Defects detected during testing will be noted, checked and corrected many times.

### *C. Structural design for Chatbot*

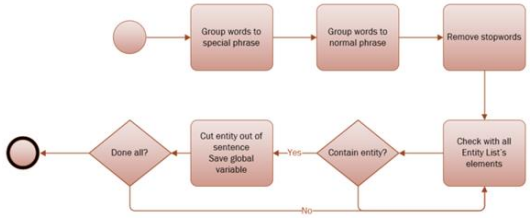
**Create training model**

Data sets in the form of excel sheets, will be conducted to build data models to be included in training for machine learning. The first is to create a bag of word based on a list of words after pretreatment, which will be used to vectorise the sentences in each topic. After vectoring the sentences, turn into a training file and use the SGDClassifier tool of the SKLearn library and apply the SVM model to train the model of a subject recognition machine.

*Building process of user input processing*

When starting a session, the chatbot will receive the input data to conduct processing before being included in the model. The input data will be processed through the following steps:

• Remove icon symbols, emoji
• Give up words that indicate laughing
• Correct common misspelled words
• Replace abbreviations with full form
• Skip nonsense repeated characters in a word
• Remove punctuation and special characters
• Pair complex words and special phrases



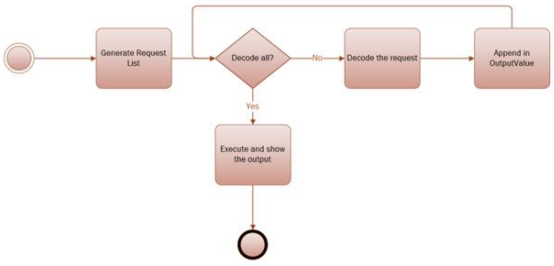

**Processing and giving results**

After the preprocessing of raw text from user input, the output of the process will be standardized text and can be entered into the algorithm model to identify the speaker's intent. In the processing to produce results, the system will identify the topic and content the user is talking about, then map with the answer set to produce the results as well as call the appropriate API or functions, depending upon the request from the user input command.

•Step 1: The model will identify the result of the preprocess and proceed to identify the topic group.

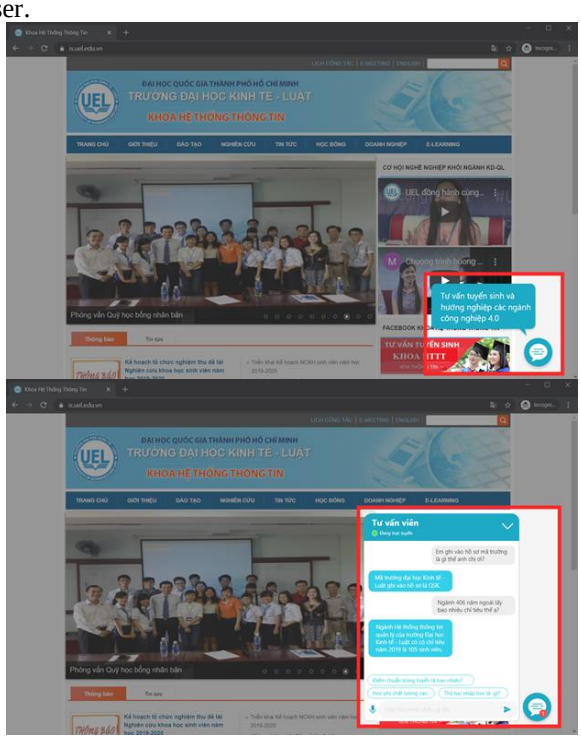
•Step 2: Identify the content that users want to talk about by loading the machine learning training model created in the model building section.

•Step 3: Once you've determined the speaker's intent, get the IntentID from the result of the function that runs the identification model, we'll proceed with processing to issue a set of requests for the response process.

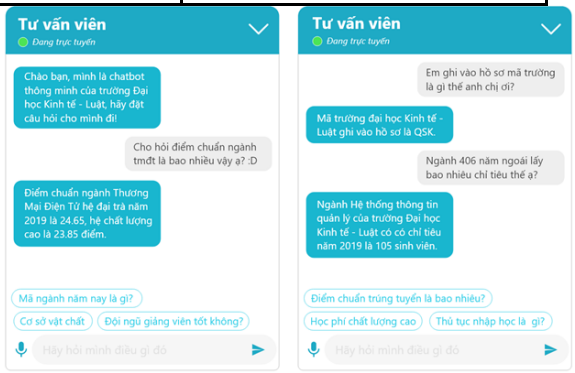



### *D. Interface design for Chatbot*

The research team designed and built a dialog interface with chatbot based on proximity, friendly and easy to use, with the structure of simple functional areas and few symbols, need to focus on the internal container communication area between chatbot and user.


| Number of questions to train: ***665***<br>Number of questions to test: ***396*** | | | | | |
|---|---|---|---|---|---|
| Times to check | 1 | 2 | 3 | 4 | 5 |
| Correct identification | 381 | 376 | 372 | 374 | 370 |
| Ratio | 96.2 | 95 | 94 | 94.4 | 93.4 |
| Average ratio | 94.6 % | | | | |



| CASE Times test | Test A | Test B | Test C | Test D | Test E |
|---|---|---|---|---|---|
| | The number of correctly identified sentences is intent on the 100-sentence set | | | | |
| 1st | 95 | 88 | 83 | 78 | 68 |
| 2nd | 96 | 87 | 85 | 76 | 65 |
| 3rd | 94 | 89 | 82 | 77 | 67 |
| 4th | 94 | 85 | 83 | 75 | 64 |
| 5th | 91 | 86 | 84 | 76 | 62 |
| 6th | 95 | 87 | 83 | 77 | 64 |
| 7th | 93 | 84 | 81 | 76 | 66 |
| Average | 94 | 86.6 | 83 | 76.4 | 65.1 |

## IV. TEST, EXPERIMENT AND DEPLOYMENT

### *A. Reassess the accuracy of the model*

**The Evaluate 1:** Performance of the model

Conduct a training with 5 random questions, then will take the remaining questions to test the ability to identify the topic of the model.

After retesting the model, the team authors concluded that the model's performance is about 94.6%. The accuracy rate is perceived to be high, partly because the similarity between the similar questions is quite high.

**Evaluate 2:** Field experiments

Conducting a test on 100 random questions belonging to the dataset, in each case the questions will be changed data and put into the identification system to check how many correctly identified questions remain.

A: Raw Question Question. B:Randomly drop 10% of words in sentences. C:Randomly skips 15% of the words in the sentence. D: Randomly removes 20% of the words in the sentence. E: Randomly remove 50% of the sentences of the sentence.

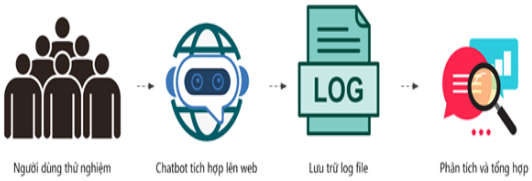
The accuracy of the model is 94%. In case the comb loses about 10% of the words in the sentence, the accuracy of the model is about 86.6%, 83% when the comb loses 15% and 76.4% when the comb randomly loses 20% of words.

When omitting half of the keywords in the sentence, the accuracy is below the average, about 65.1%. So when users ask questions that are not too misleading in the data set, the accuracy of the algorithm is relatively high - over 80%. However, the more mistakes or omissions of words, the more the accuracy will decrease, only about 65-75%.

*How to fix and improve:* To increase the accuracy of the identification model, we can enrich the question data set to increase the coverage of cases.

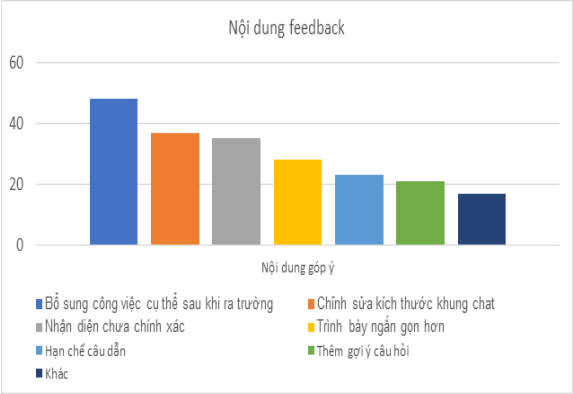
### *B. Experiment*

After completing the first version of the model and the interface, the team proceeded to deploy and integrate on the website system to experiment and collect real data of users. After collecting, we will synthesize, edit and update the version, then conduct a test run to collect and verify the next time.




After summarizing, arranging and grouping the data obtained from experiments, the team conducted:

- Edit 11 current content still encounter errors.
- Additional 22 new content and topics.
- Record 13 unsatisfied issues about the web system and presentation format.
- Recorded 7 suggestions for improvements to the interface, content, scripts, ...




## V. CONCLUSION AND RECOMMENDATIONS

### *A. Conclusion*

The thesis "Building AI Chatbot to support admissions and career guidance for universities" has been explored, analyzed and understood the difficulties, shortcomings and problems arising in career counseling, enrollment support. Besides, the thesis has successfully built a structured dataset about enrollment orientation and apply of natural language processing, machine learning to build identification models. The thesis will contribute to overcome and improve the performance of university admissions consulting.

### *B. Limitation*

During the research and implementation of the project, the authors realized a few shortcomings and limitations as follows:

• Questionnaire survey is not really optimal and there are many shortcomings in the data collection process. Therefore, it leads to redundant data collection, which does not serve the research process well.

• The survey scope is not wide enough and the number of samples collected is limited.

• The system has not been fully developed and integrated into the messenger platform and message box of the university website.

• The data set of questions and answers is collected only for a short time and not diverse enough in the data source, and so the data still has many shortcomings and does not cover all cases of questions from students as well as parents.

### *C. s Recommendations*

• Firstly, integrating with Speech To Text speech processing system to support advice by receiving user voice commands or chatting online chat with chatbot.

• Secondly, Integrating chatbot into university website system, mailbox system of the fanpage, advisory mail system, hotline automatic answering system.

• Thirdly, expanding the chatbot support topics: Integrated into student handbook, postgraduate training advice, master's degree consulting.

• Fourthly, building and developing chatbot as a smart teaching assistant to support teachers in the teaching process and answer questions about frequently asked questions about document management, lecture slides, support for submission, software installation,

• Fifthly, developing virtual assistant application for student life, integrating for paying tuition, buying curriculum and paying for cafeteria meals, bus station lookup and library booking.

• Finally, Chatbot application to advise secondary students in selecting specialized classes for high schools.

## REFERENCES

[1] Weiyu, W., & Keng, S. (2018). Living with Artificial Intelligence-Developing a Theory on Trust in Health Chatbots.
[2] Massimiliano, Katarzyna, Federica, T., & Carlo, M. (2018). Chatbot in a Campus Environment: Design of LiSA, a Virtual Assistant to Help Students in Their University Life. Human-Computer Interaction.
[3] Ismail, M. & Abejide, A. I. (2019). Lecturer-Chatbot: An AI for Advising Struggling Students in Introductory Programming. SACLA 2019.
[4] Bii, P., Too, J., & LangatAn, R. (2013). Investigation of student's attitude towards use of chatbot technology in Instruction: the case of Knowie in a selected high school. Educational Research.
[5] Lai, T. H. Y., Trieu, H. A., Nguyen, T. T. (2018). Research on building a chatbot system to support advising VMU students. Vietnam Maritime University.
[6] Gayatri, N., & Johnson, S. (2018). Chatbot as a Personal Assistant. International Journal of Applied Engineering Research.
[7] Vijayarani, S., Ilamathi, J., & Nithya (2015). Preprocessing Techniques for Text Mining.
[8] Le, H. S., Lee, J. H., & Lee H. K. (2017). Analyzing visitors' preferences on tourism accommodation services by opinion mining. The Journal of Internet Electronic Commerce Research, 113.
[9] Das, B., & Chakraborty S. (2018). An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation
[10] Grzegorczyk, K. (2018).Vector representations of text data in deep learning.

[11] Yuwono, S. K., Biao, W., & D'Haro, L. F. (2018). Automated scoring of chatbot responses in conversational dialogue
[12] Radziwill, N., & Benton, M. (2017). Evaluating Quality of Chatbots and Intelligent Conversational Agents.
[13] Leuthäuser M. (2010). Object-oriented programming with Python.
[14] Le, H. S., & Lee H. (2015). A Competitive Analysis of Coffee Franchise Brands on Social Media Using Big Data. The Journal of Internet Electronic Commerce Research,, 15(5), 17-29.