# Deep Learning for Conversational Agent via Context Question Answering Model

**Pui Huang Leong[1], Ong Sing Goh[2], Yogan Jaya Kumar[3]**

[1]Faculty of Computing and Information Technology, Tunku Abdul Rahman University College (TAR UC),
Malaysia, leongph@tarc.edu.my

[2]Centre for Advanced Computing Technology, Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka (UTeM), Malaysia, goh@utem.edu.my

[3]Centre for Advanced Computing Technology, Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka (UTeM), Malaysia, yogan@utem.edu.my

## ABSTRACT

The usage of a conversational device communicate knowledge on behalf of humans has attracted popularity and has been widely employed. As a consequence, this study seeks to analyze and improve the current business platform by providing the ability to offer immediate feedback by integrating Deep Learning to the conversational agent. Our research focuses on Artificial Intelligence (AI) to include Natural Language Processing (NLP) by offering a paradigm focusing on Deep Learning via Context Question Answering Model. This model enables the conversational agent to make wiser and stronger choices to be able to understand and generate their own answers to unknown queries. The testing will be performed in the quantitative approach using Black-box Testing and User Experience (UX) Testing. The Black-box Testing findings show that the average time taken for Deep Learning to produce a response is approximately two seconds. As far as accuracy is concerned, the conversational agent is capable of attaining the accuracy of 67.65 percent. The accuracy could be improved from time to time by providing more training to the model. Apart from this, the test results from User Experience Testing indicates that 84.3 percent of the respondents are satisfied with the conversational agent, whereas 86 percent of respondents would use this conversational agent again. The experiments provide positive outcomes and evidence of the progress of this research, with tremendous contributions to the area of artificial conversational agent.

**Key words :** Artificial Intelligence, Conversational Agent, Context Question Answering Model, Deep Learning.

## 1. INTRODUCTION

The application of artificial conversational agents to communicate information on behalf of humans has been a major trend that is being utilized intensively, especially in the field of commerce [1], [2], [3], [4], [5], entertainment [6], tourism [7], medical [8], [9], [10], [11], [12], [13] as well as education [14], [15], [16], [17], [18]. According to the author in [19], by implementing AI, the banking industry could enjoy possible cost savings of up to 25 percent in Information Technology operations. The cost saving is primarily due to AI's ability to provide innovative smart technology when helping managers to make decisions. It is due to factors such as resource efficiency, cost-effectiveness, knowledge availability and the ubiquitous of information. Conversational agent, or very often address as chatbot, refers to the attempts of the computer system to imitate human-machine communications. Besides, the purpose of this research is to automate the process of minimising the need for human intervention, explicitly for the botmaster to conduct information maintenance manually. The majority of existing approaches optimise the management of the knowledge base by a supervised approach whereby the botmaster needs continuously supply the necessary details and train the bot, which is inefficient and time-consuming. The conversational agent in this study is further advanced to ensure smarter decision-making by delivering fast and precise knowledge via the emergence of Deep Learning. Deep Learning allows a conversational agent to derive self-understanding answers on its own. Sushi King is chosen to be implemented as the domain-specific knowledge in this study. In this study, T.A.R.I.E. (The AI Revolutionisation in Embodied Conversational Agent) is the conversational agent with the role of customer service specialist to respond to users' prompt queries.
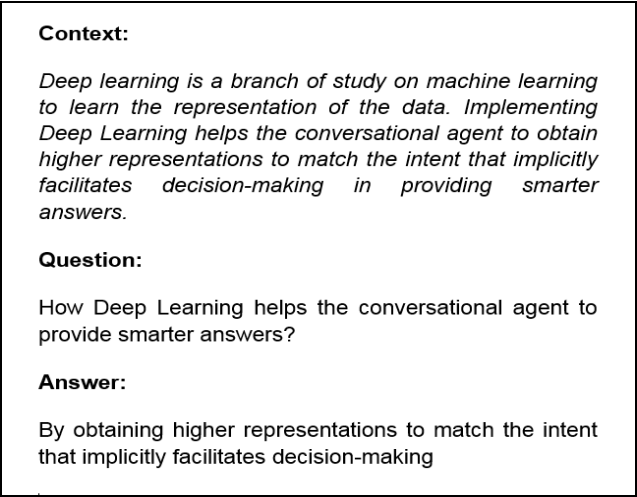
## 2. DEEP LEARNING

Deep Learning is a division of machine learning to analyses and interpret the information. Deep Learning has been applied in several areas. For instance, the work in [23] used Deep Learning in translating texts whereas the work in [24] used Deep Learning for gesture recognition. Implementing Deep Learning lets the conversational agent achieve higher representations to suit the purpose that indirectly enables decision-making to deliver smarter answers. The

conversational agent is far more advanced as it can produce responses on its own, without depending on the preset list of questions and answers. Nonetheless, since this model is smart to make its own decisions and to produce answers based on its comprehension from scratch, the probability of responding to grammatical and spelling errors must be taken into account. This work relies on Deep Learning to seek answers to any unanswered questions to make smarter decisions. When questions are not identified within common knowledge or Frequently Asked Questions (FAQs) knowledge bases, Deep Learning is triggered to look for possible answers from the corporate portal. This approach is able to prevent any random-answer generation or answer-redirecting from the conventional way for any unanswered queries. As this process is achieved automatically, manual supervision on the conversational agent could be reduced. Before the Deep Learning works, several configurations would need to be completed and will be addressed further in the following section.

## 2.1 Context Question Answering Model

Context Question Answering Model will be adapted to the Deep Learning of the conversational agent, T.A.R.I.E in this study. This model operates for a defined context to seek for the solution to the request, where the response to each query is part of the context. This model is developed based on end-to-end architecture of neural networks for answering questions in the form of reading comprehension [20]. The illustration of the model with Context, Question and Answer is portrayed as in Figure 1. This model suits the research objective of this study to exploit the use of Deep Learning in generating its own responses whenever unforeseen circumstances arise. Deep Learning will then be triggered to search for the answer through the corporate portal for unknown queries. In such a case, the corporate portal serves as the context for this model.

> **Context:**
>
> *Deep learning is a branch of study on machine learning to learn the representation of the data. Implementing Deep Learning helps the conversational agent to obtain higher representations to match the intent that implicitly facilitates decision-making in providing smarter answers.*
>
> **Question:**
>
> How Deep Learning helps the conversational agent to provide smarter answers?
>
> **Answer:**
>
> By obtaining higher representations to match the intent that implicitly facilitates decision-making

**Figure 1:** Overview of Question Answering Model

The adoption of this model enables the conversational agent to understand the background based on the query prompt by the user and generate a response accordingly. In order for the

conversational agent to learn how to generate the response based on the understanding from the context, training is mandatory.

Thus, Context Question Answering Model is needed to be trained by providing the dataset which matches with its format. Stanford Question Answering Dataset (SQuAD 2.0) dataset will be used to train the model and will be addressed further in the next Section.

## 2.2 SQuAD 2.0 Dataset for Context Question Answering Model

The dataset, SQuAD 2.0, which follow the Context Question Answering format is needed to train the model. SQuAD 2.0 is a dataset created specifically for reading comprehension, derives from the queries presented by the community on Wikipedia pages where the solution to each question is a portion of the document. SQuAD2.0 is an open-source that includes more than one hundred thousand simulations dataset for model [21]. The dataset is open source and has been deployed by many other researchers; therefore, the SQuAD 2.0 datasets will be used as the training dataset for the Context Question Answering Model in this research. As Sushi King is chosen to be the target corporate in this research, the model is then trained with Sushi King datasets to ensure smarter learning and decision for domain-specific knowledge. Before using the model, all the required packages are needed to be installed running the command as this serves as the prerequisites before the model could be used.

## 2.3 SQuAD 2.0 Dataset Merged with Sushi King Dataset

As mentioned in Introduction, Sushi King is chosen to be implemented as the domain-specific knowledge in this study. In order to come out with a smarter conversational agent, the model is then trained with the dataset from the target corporate, Sushi King. The information from Sushi King portal is extracted into JavaScript Object Notation (JSON) file before merging into the SQuAD 2.0 datasets to be used as the improvise datasets to train the Context Question Answering Model. The overall timeline to train the model is portrayed as in Figure 2. The Sushi King information is extracted into JSON file for the compatibility of the training dataset with the entire duration of two days. After the Sushi King dataset is created, the dataset is then merged into SQuAD 2.0 dataset. Both the files are in JSON format, and the entire duration for the merging is approximately six days. The model is then trained with Google Colab with the new dataset, which took approximately one month, twenty-seven days to complete. Google Colab is an online browser-based application that helps researchers to train Deep Learning models on the computer at no cost. This is a free cloud provider, which provides free GPU to train models with the massive dataset. This application is great to enhance the Python coding that works with Deep Learning libraries such as Keras and TensorFlow. The most significant attribute differentiating Google Colab from many other online cloud providers is that it offers GPU and is completely at not chargeable.
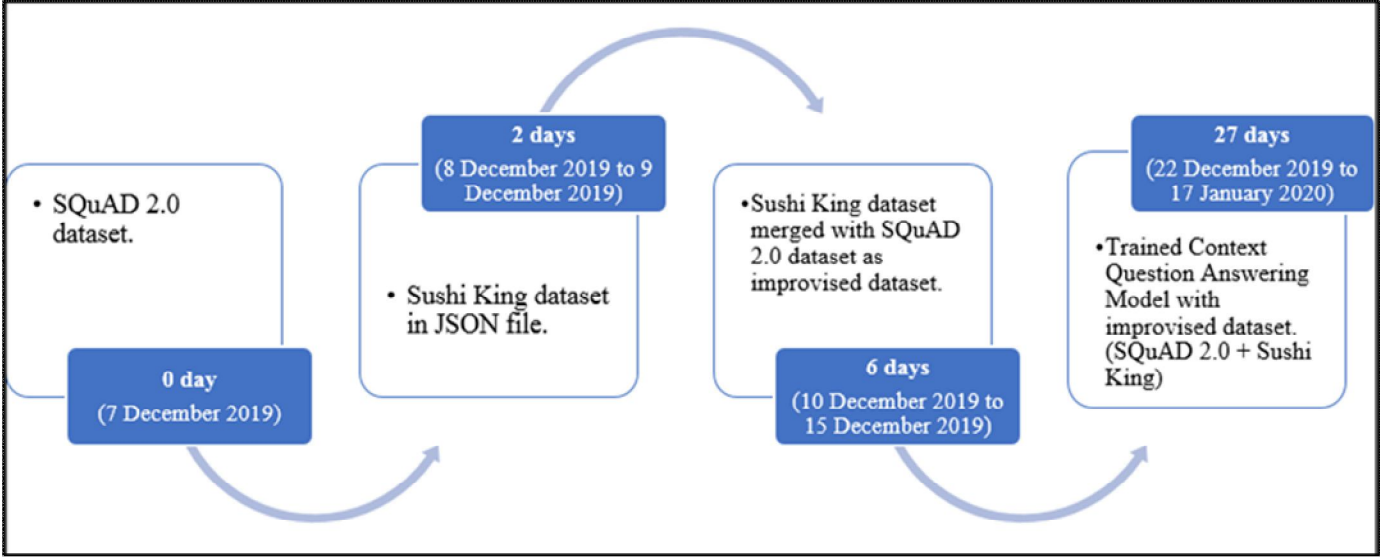
**Figure 2:** Timeline to Train the Model

| Step | Main process | Process details |
|---|---|---|
| 1 | *Context extraction* | The context of the selected corporate web pages will be extracted. |
| 2 | *Question and context encoder* | Compute the question and context encoder. The query will be converted to word-level and character-level embedding in order to generate new representation. The new representation is needed to match with the entire context in the next stage. |
| 3 | *Gated attention-based recurrent network* | Compute gated attention-based recurrent network. The new representation will be matched to the entire context to form question and context representation. This is crucial to indicate the relevance between the passage parts to the question. |
| 4 | *Self-matching attention* | Compute self-matching attention because even though the gated attention is used to mark down the important keywords in the previous stage to determine question-aware passage, the knowledge of the context might still be limited. In order to solve this, the question-aware passage is matched again with the entire context. |
| 5 | *Output layer (pointer networks)* | Compute pointer networks. The words in relation to the question have been determined. However, the sequence or position of the words might not be in the position. |
| 6 | *Display answer* | An answer will be generated based on Deep Learning interpretation. |

**Figure 3:** Algorithm for T.A.R.I.E Deep Learning

## 2.4 Deep Learning for Conversational Agent (T.A.R.I.E.)

The concept of Deep Learning searching for plausible information from the corporate platform is drawn from the excellent research of Microsoft Asia team [20]. The Context Question Answering Model is developed based on R-NET, an end-to-end architecture of neural networks for answering questions in the form of reading comprehension. This model is structured to respond to a query based on the given context.

This model has been implemented by many other researchers for research concerning reading comprehension. As this approach requires enormous datasets for training, SQuAD 2.0 is regarded as the most common and popular training datasets. This research involves answer lookup by Deep Learning to look for plausible information from the corporate portal and to generate own responses. Context Question Answering Model with R-NET, end-to-end Neural Networks is regarded as the best model to be exploited. This model works for a defined context to provide a response to the query, where a portion of the context is the response to each query. This model initially matches the query and context with gated attention-based recurrent networks to achieve a description of the problem-aware context. Then, self-matching attention to optimizing the representation by matching the context against itself that essentially encodes knowledge from the entire passage is performed. Pointer networks will then be used to determine the positions of the responses.

In order to build a smarter model, Deep Learning in this study is trained via SQuAD 2.0 and the Sushi King dataset. Figure 3 shows the detailed algorithms of Deep Learning implementation for T.A.R.I.E. whereas Figure 4 shows the flow of Deep Learning implementation for T.A.R.I.E. The entire Deep Learning process is achieved automatically, which indirectly reduce human intervention. This approach can address the conventional approach to provide random-answer generation or answer-redirecting whenever unforeseen circumstances arise, which constituent to the users doubting the intelligent of the conversational agent. Figure 5 shows the sample conversations where the responses are generated with Deep Learning, whereas Figure 6 shows the sample conversation where the responses are generated without Deep Learning.

Deep Learning is triggered whenever unforeseen circumstances arise. This situation allows Deep Learning to render representations that match the intent of the corporate portal. Through this model, the conversational agent is capable of making its own decisions based on the contextual generalization and generate an appropriate response to the user.
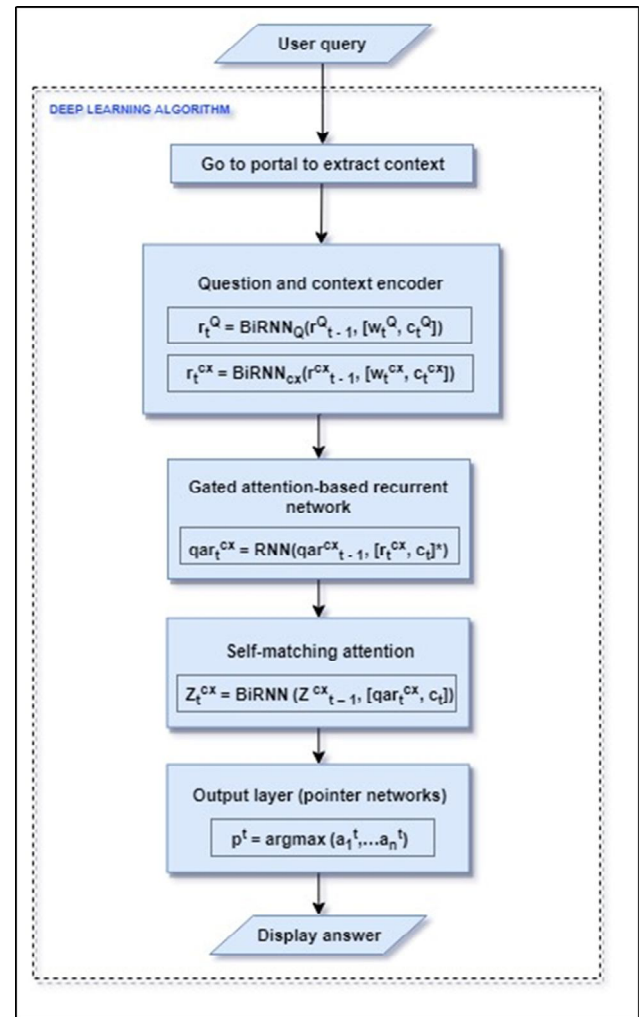


**Figure 4:** The Flow of the Deep Learning Implementation for T.A.R.I.E.
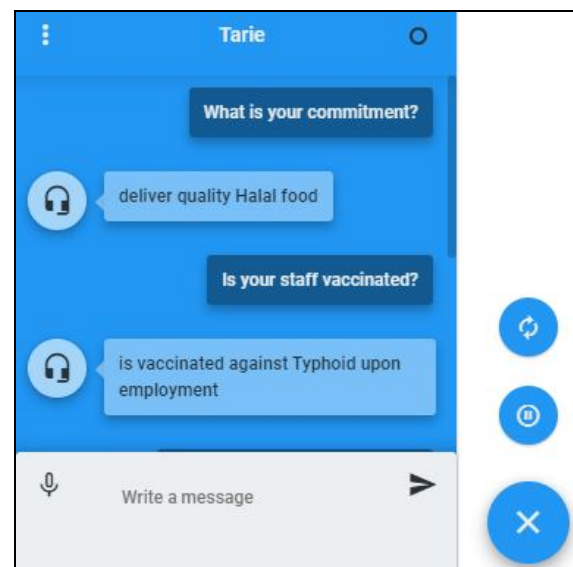


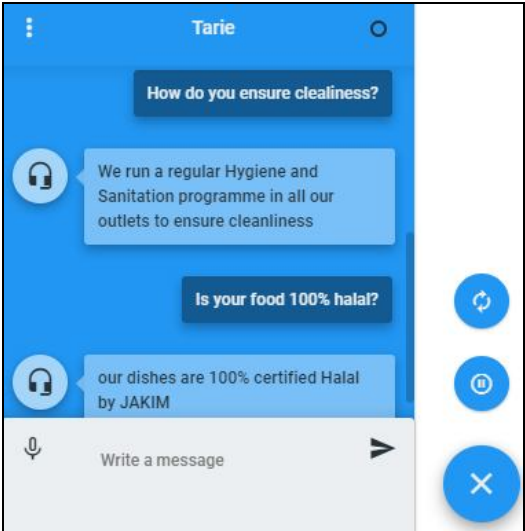**Figure 5:** Responses Generated with Deep Learning (Continue)

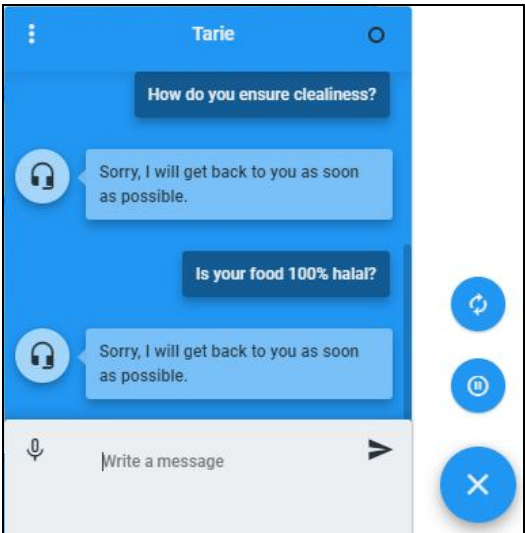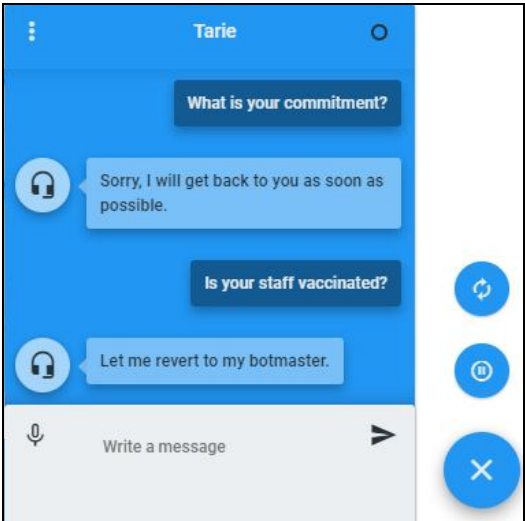**Figure 5:** Responses Generated with Deep Learning





**Figure 6:** Responses Generated without Deep Learning

## 3. TESTING AND ANALYSIS

In order to perform testing for the speed (response time), and accuracy for Deep Learning, 34 questions adapted from Sushi King customers, will be used. The testing is performed in a guided environment in which the customers are given the web pages of Sushi King. Based on the web pages, the Sushi King customers are required to ask queries pertaining to Sushi King. In order to assure that Deep Learning works as intended, the total of 34 questions generated is used as the questions for Deep Learning.

**Table 1:** Test Results for Deep Learning Questions

| Question | Speed (in milliseconds) | Accuracy Score |
|---|---|---|
| 1 | 2,744.70 | 2 |
| 2 | 2,442.16 | 1 |
| 3 | 2,364.64 | 0 |
| 4 | 2,769.60 | 0 |
| 5 | 2,585.66 | 2 |
| 6 | 2,460.49 | 0 |
| 7 | 2,463.50 | 1 |
| 8 | 1,912.32 | 2 |
| 9 | 2,417.62 | 2 |
| 10 | 2,514.28 | 2 |
| 11 | 2,723.21 | 2 |
| 12 | 2,221.27 | 2 |
| 13 | 2,500.96 | 2 |
| 14 | 1,946.13 | 2 |
| 15 | 3,115.97 | 1 |
| 16 | 2,693.89 | 2 |
| 17 | 1,974.37 | 2 |
| 18 | 1,973.52 | 0 |
| 19 | 2,370.13 | 1 |
| 20 | 2,661.25 | 2 |
| 21 | 2,467.85 | 1 |
| 22 | 1,924.98 | 2 |
| 23 | 2091.78 | 0 |
| 24 | 1,901.35 | 2 |
| 25 | 2,358.58 | 2 |
| 26 | 1,891.77 | 2 |
| 27 | 2,054.27 | 1 |
| 28 | 2,558.41 | 2 |
| 29 | 2,520.10 | 2 |
| 30 | 1,928.30 | 0 |
| 31 | 1,978.87 | 0 |
| 32 | 2,071.49 | 0 |
| 33 | 2,277.71 | 2 |
| 34 | 2,045.13 | 2 |
| **Average** | 2321.36 | [(2x20) + (1x6) + (0x8)] / (34 x 2) x 100% = 67.65% |

Based on the domain-specific queries, Deep Learning will then perform answer lookup from the Sushi King portal and generate responses accordingly. Deep Learning will perform reasoning and understanding and come out with its own responses. The scoring for accuracy parameter is adapted based on the Loebner Competition [22]. Table 1 portrays the test results for Deep Learning questions. Figure 7 and Figure 8 portray the test results in the form of graph.
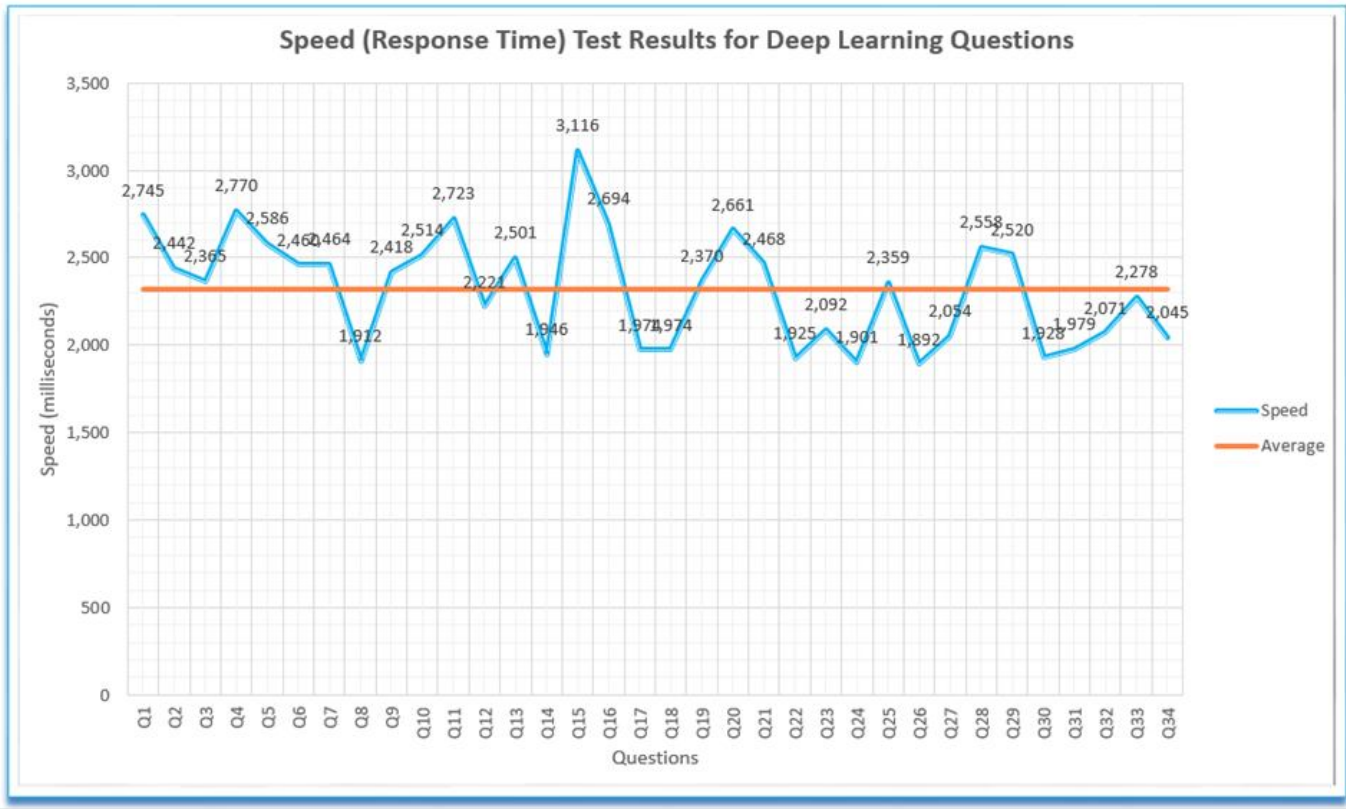
**Figure 7:** Speed (Response Time) Test Results for Deep Learning
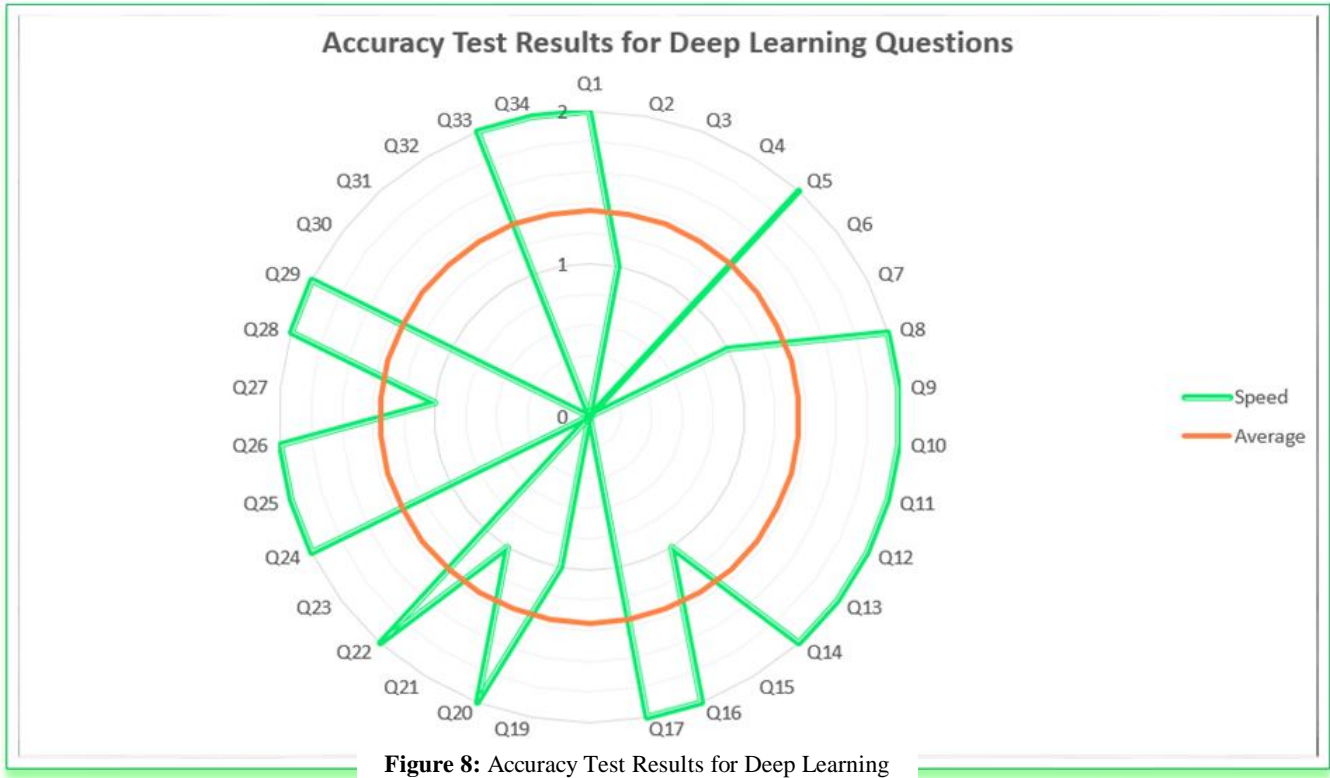


**Figure 8:** Accuracy Test Results for Deep Learning

The analysis from Table 1 demonstrates that the average speed, in terms of response time for the total of 34 questions is 2321.36 milliseconds (2.321 seconds). This shows that the chatbot is capable of responding in less than three seconds. The shortest time recorded, 1,891.77 milliseconds is retrieved from Questions 26, whereas the highest time recorded, 3,115.97 milliseconds is retrieved from Questions 15. The analysis shows that the time taken for the chatbot to retrieve the answer is depending on the chatbot to perform the Deep Learning algorithm. The outcome indicates that the conversational agent is capable to perform answer lookup through Deep Learning. The analysis from Table 1 indicates responses generated by Deep Learning achieved 67.65 percent accuracy. In the total of 34 questions, 20 questions are able to provide the correct responses, whereas 6 questions are able to provide partially correct responses. The chatbot is unable to offer accurate responses to 8 questions. The issue pertaining to the inaccuracy of Deep Learning could be improved via training. All the responses generated by Deep Learning will be saved into the repository for knowledge maintenance. The result with 67.65 percent accuracy shows that Deep Learning is able to provide accurate responses promptly and the accuracy could be improved from time to time via training. The conventional method required botmaster to administer knowledge maintenance manually whenever the conversational agent encounters unknown requests, which is time-consuming and tiresome. Deep Learning could solve the conventional approach whereby the process is achieved automatically, approximately three seconds without the interference of botmaster.

## 4. CONCLUSION

Generally, the results from the testing indicate the success of the study. Deep Learning will be used to locate information from the corporate portal whenever unforeseen circumstances arise. Even though Deep Learning takes approximately 3 seconds to response, the process is achieved automatically without human intervention. The results from the testing show that the conversational agent is capable of selecting the best model to undergoes the series of answer lookup. The proposed model is able to solve the conventional approach to reduce human interference due to the automated process. Furthermore, the analysis from User Experience Testing stipulates that most respondents are pleased with the conversational agent and will use it again. The respondents claimed that the chatbot is useful and they are able to retrieve answer quickly and effectively via this chatbot.

## ACKNOWLEDGEMENT

## REFERENCES

[1]. Ko, M. and Lin, Z., 2018. **Chatbot: A Chatbot for Business Card Management**. IUI '18 Companion: Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, 1-2.

[2]. Choi, H., Hamanaka, T., and Matsui, K. (2017). **Design and Implementation of Interactive Product Manual System using Chatbot and Sensed Data**. 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE 2017), pp. 1-5.

[3]. D'silva, G.M., Thakare, S., More, S., and Kuriakose, J., 2017. **Real World Smart Chatbot for Customer Care using a Software as a Service (SaaS) Architecture**. International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), pp. 658–664.

[4]. Bhawiyuga, A., Fauzi, A., Pramukantoro, E.S. and Yahya, W. (2017). **Design of E-Commerce Chat Robot for Automatically Answering Customer Question**. 2017 International Conference on Sustainable Information Engineering and Technology (SIET), pp. 159-162.

[5]. Thomas, N.T. (2016). **An E-business Chatbot using AIML and LSA**. 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2740-2742.

[6]. Cuayahuitl, H. (2017). **Deep Reinforcement Learning for Conversational Robots Playing Games**. 2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids), pp.771–776.

[7]. Argal, A., Gupta, S., Modi, A., Pandey, P., Shim, S., and Choo, C. (2018). **Intelligent Travel Chatbot for Predictive Recommendation in Echo Platform**. 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), pp.176–183.

[8]. Stal, S., Kramer, L.L., Tabak, M., Akker, H., and Hermens, H. (2020). **Design Features of Embodied Conversational Agents in eHealth: A Literature Review**, International Journal of Human-Computer Studies, 138: 1-22.

[9]. Lee, D., Oh, K. and Choi, H. (2017). **The ChatBot Feels You – A Counseling Service Using Emotional Response Generation**. 2017 IEEE International Conference on Big Data and Smart Computing. (BigComp), pp. 437-440.

[10]. Madhu, D., Neeraj, J.C.J, Sebastain, E., Shaji, S., and Ajayakumar, A. (2017). **A Novel Approach for Medical Assistance Using Trained Chatbot**. International Conference on Inventive Communication and Computational Technologies, pp. 243–246.

[11]. Oh, K., Lee, B.K., and Choi, H. (2017). **A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation**. 2017 18th IEEE International Conference on Mobile Data Management (MDM), pp. 371-375.

[12]. Ghosh, A., Chaki, R. and Chaki, N. (2016). **Checkpoint Based Multi-Version Concurrency Control Mechanism for Remote Healthcare System**. 2016 International

Conference on Advances in Computing, Communications and Informatics (ICACCI), pp.382–389.

[13]. Sankarananrayanan, S., and Wani, S.M.A. (2014). **NFC Enabled Intelligent Hospital Appointment and Medication Scheduling**. 2nd International Conference on Information and Communication Technology, pp.24–29.

[14]. Song, D., Oh, E.Y., and Rice, M. (2017). **Interacting with a Conversational Agent System for Educational Purposes in Online Courses**. 2017 10th International Conference on Human System Interactions (HSI), pp. 78-82.

[15]. Ranoliya, B.R., Raghuwanshi, N. and Singh, S. (2017). **Chatbot for University Related FAQs**., 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI),  pp. 1525-1530.

[16]. Fadhil, A. and Villafiorit, A. (2017). **An Adaptive Learning with Gamification and Conversational UIs: The Rise of CiboPoliBot**. 25th User Modeling, Adaption, and Personalisation, pp.408–412.

[17]. Kumar, M.N., Chandar, L.P.C., Prasad, V.A., and Sumangali, K. (2016). **Android Based Educational Chatbot for Visually Impaired People**. 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-4.

[18]. Benotti, L., Martínez, M.C. and Schapachnik, F. (2014). **Engaging High School Students Using Chatbots**. 19th Annual Conference on Innovation and Technology in Computer Science Educatioin (ITICSE), pp.63–68.

[19]. Accenture, (2018). **Redefine Banking with Artificial Intelligence**. [online] Available at:

https://www.accenture.com/acnmedia/pdf-68/accenture-rede fine-banking.pdf [Accessed on 17 September 2019].

[20]. Natural Language Computing Group, Microsoft Research Asia, (2017). **R-NET: Machine Reading Comprehension with Self-Matching Networks**, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 189-198.

[21]. SQuAD Stanford NLP Group (2018). SQuAD2.0 **The Stanford Question Answering Dataset**, [online] Availble at: https://rajpurkar.github.io/SQuAD-explorer/ [Access 26 July 2019].

[22]. Loebner Prize., 2018. **Turing Test 2018: Results**, [online] Available at:

https://artistdetective.wordpress.com/category/loebner-prize/ [Accessed on 7 September 2018].

[23] A. V. Zadgaonkar and R. S. Vairagade. **An Approach for Translation of Text in Images using Deep Learning Techniques**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9, pp. 808-812, February 2020.

[24] C. Ng and A. Chua. **Training of a Deep Learning Algorithm for Quadcopter Gesture Recognition**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9, pp. 211-216, February 2020.