The Evaluation of User Experience Testing for Retrieval-based Model and Deep Learning Conversational Agent

Pui Huang Leong[1], Ong Sing Goh[2,] Yogan Jaya Kumar[3], Yet Huat Sam[4], Cheng Weng Fong[5]
Faculty of Computing and Information Technology, Tunku Abdul Rahman University College (TAR UC), Johor, Malaysia[1, 5]
Centre for Advanced Computing Technology, Faculty of Information and Communication Technology[2, 3]
Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia[2, 3]
Faculty of Accountancy, Finance and Business, Tunku Abdul Rahman University College (TAR UC), Johor, Malaysia[4]

***Abstract*—The use of a conversational agent to relay information on behalf of individuals has gained worldwide acceptance. The conversational agent in this study was developed using Retrieval-based Model and Deep Learning to enhance the user experience. Nevertheless, the successfulness of the conversational agent could only be determined upon the evaluation. Thus, the testing was performed in the quantitative approach via questionnaire survey to capture user experience upon the usage of the conversational agent in terms of Usability, Usefulness and Satisfaction. The questionnaire survey was tested via statistical tool for reliability and validation test and proven to be carried out. The test results indicate positive experience towards the usage of the conversational agent and the outcome of the testing showed promising results and proof the success of this study, with immense contributions to the field of conversational agent.**

***Keywords*—Conversational agent; retrieval-based model; deep learning; user experience testing; usability; usefulness; satisfaction**

I. Introduction

User experience may be unique to the extent that it affects human interpretation and feeling regarding a service or program. In brief, user experience was about how a consumer communicates with a device and its interactions. This broadness means that the user experience included several facets, when documenting user interaction while utilising a conversational agent. User Experience testing was conducted to capture user experience upon interacting with the conversation agent. By the end of the testing, questionnaire survey was handed out to obtain the feedback and satisfaction towards the usage of the conversational agent.

The international agreement on the ergonomics of contact between human beings as stated in the ISO 9241-210:2010 depicted the User Experience as the expectations and reactions of an individual arising from the usage or expectation of service, device or system. User experience involved the feelings, desires, attitudes, physical and psychological reactions, habits of all users that emerge before, during and after the usage. According to [1], when measuring user experience while using a chatbot, user experience could be separated into three specific needs, namely Usefulness, Usability and Satisfaction.

User experience testing was a process in which the interface and the chatbot features were verified by end users who execute specific tasks under practical environments. This test aimed to assess the user experience in terms of Usability, Usefulness and Satisfaction of the conversational agent and to ascertain if the application was functional. At the end of the testing, a survey was carried out via Google Form to gather user' response and satisfaction towards the usage of the conversational agent. Prior to the released of the questionnaire survey for User Experience Testing, validation of the research instruments will be conducted to ensure the survey was ready to be escalated for the real testing.

The next section discussed the three aspects of the User Experience Testing in details. Section III discussed the research instruments questionnaires followed by Section IV to discuss the validations of the research instruments. Next, sample size population was discussed to provide insight on how the total number of respondents were determined. Section VI discussed the testing and analysis followed by the last section to discuss the conclusion of the study.

II. User Experience Testing

Generally, the user experience was measured in three aspects, namely Usefulness, Usability and Satisfaction via the quantitative method in this research. The user experience testing was scoped into three measures as illustrated in Fig. 1.

The following section discussed the Usability measure, Usefulness measure and Satisfaction measure of the User Experience Testing in depth.

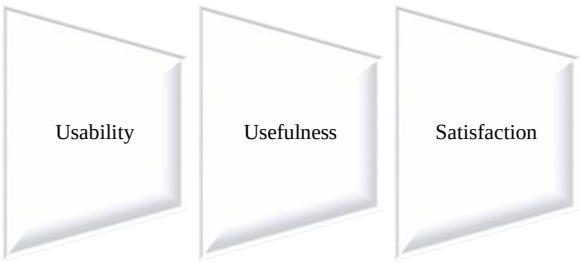

Fig. 1. User Experience (UX) testing [1].

A. Usability Measure

Usability was part of the broader phrase "user experience" which refers to a software or service was readily viewed or utilised. The international standard ISO 9241-11:2018 concept of usability was: "the degree to which a service or product may be used by specified users to accomplish defined goals with effectiveness, efficiency and satisfaction in a defined context of use." The use of structured surveys was a common and cost-effective method for usability assessments. A typical usability assessment known as Usability Metric for User Experience (UMUX) as shown in Fig. 2 has been used to assess usability.

According to [2], the Usability Metric for User experience was versatile for a larger user experience variable to function as a usability element. UMUX was used in this study to evaluate the usability of user experience. The authors in [5] and [6] have been adopting UMUX to measure the usability experience of the users.

B. Usefulness Measure

Usefulness was described as being useful when it comes to quality or fact. In Technology Acceptance Model (TAM), perceived usefulness has been identified as one of the variables influencing the usage and adoption by specific users of information systems and technologies. TAM, founded by [3], was one of the most common methods of analysis to forecast the usage and recognition by specific consumers of information systems and technologies.

TAM as in Fig. 3 has been extensively examined and validated by numerous experiments that investigate the individual behaviors in acceptance of technology in diverse structures of information systems. TAM Model described there were two measures which were essential in the study of computer use behaviors, namely perceived usefulness and perceived ease of use. The author in [3] described perceived usefulness as the subjective likelihood of the prospective customer that utilising a particular application program would increase the efficiency of his or her work or existence. Perceive ease of use could be described as the degree to which the prospective consumer considers the target program to be effortless.

1.	[This system's] capabilities meet my requirements.						
	1	2	3	4	5	6	7
	Strongly Disagree						Strongly Agree
2.	Using [this system] is a frustrating experience.						
	1	2	3	4	5	6	7
	Strongly Disagree						Strongly Agree
3.	[This system] is easy to use.						
	1	2	3	4	5	6	7
	Strongly Disagree						Strongly Agree
4.	I have to spend too much time correcting things with [this system].						
	1	2	3	4	5	6	7
	Strongly Disagree						Strongly Agree

Fig. 2. Usability Metric for user Experience [2].

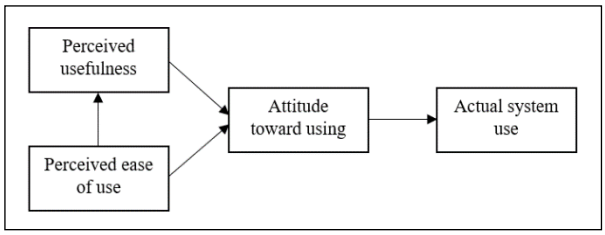

Fig. 3. Technology Acceptance Model [3].

Ease of use and perceived usefulness, according to TAM, were the most significant determinants of the actual use of the system. According to [7], TAM was amongst the essential individualistic analytical methods related to the implementation of information and communication technology (ICT). Moreover, researchers in [8], [9], [10] and [11] used TAM to test their acceptance of technologies through perceived usefulness and perceived ease of use. The relative utility of TAM was in line with the essence of this study and would be used as one of the assessing aspects of user experience.

C. Satisfaction Measure

According to the Cambridge English Dictionary, satisfaction was defined as the pleasant feeling when an individual receives something he or she wanted, or when the individual has done something he or she wanted to do. Satisfaction was one of the crucial aspects to measure user experience. The authors in [4] developed an integrated open-domain question-answering framework to test the dialogue layout that involved evaluating one of the primary factors in assessing user experience, which was the user's overall satisfaction. Two questions adapted from [4] were used to capture the overall satisfaction of the user. The next section discussed the research instruments questionnaires.

III. Research Instruments Questionnaires

Generally, the UMUX survey structure was used to measure the usability aspect. The UMUX questionnaire's benefit was that it comprised of only four reliable and relevant questions [2]. Moreover, the usefulness questions would be derived from the TAM questionnaire. The TAM questionnaire has the benefits of being a structured and widely employed test for measuring usefulness and ease of use [3]. This research adopting the questions to access usefulness. Finally, the questions to measure satisfaction aspect was retrieved from [4]. Table I illustrated the summary of the User Experience Testing questionnaires in this study.

There was a total of 16 questions in the questionnaire. The first part of the questions was asking user pertaining to the demographic to capture the age, education, occupation as well as if the user has ever used a chatbot. The next twelve questions derived from the questions from the three aspects to measure user experience upon the usage of the chatbot, namely the Usability, Usefulness and Satisfaction. These questions measured overall user experience upon the usage of the chatbot. All these questions were closed-ended type, and the data measure type was quantitative as the questions were prompted via the Likert-scale, ranging from strongly disagree (score-1) to strongly agree (score-5) which could be quantified

using numbers. The next section discussed the validation of the research instruments prior to the release of the User Experience Testing. Table I showed the summary of the User Experience Testing Questionnaires.

TABLE I. USER EXPERIENCE TESTING QUESTIONNAIRES

Measure	Questions	Source
Demographic	• What was your age? • What was your highest education? • What was your occupation? • Have you ever used a chatbot before?	[1]
Usability	• The chatbot's capability meet my requirements. • Using this chatbot was a frustrating experience. • This chatbot was easy to use. • I have to spend too much time correcting things with this chatbot.	[2]
Usefulness	• Because of this chatbot, I could quickly execute the task (retrieve answer). • This chatbot makes it hard to execute the task (retrieve answer). • Because of this chatbot, I could effectively execute the task (retrieve answer). • This chatbot was useless.	[1], [3]
Satisfaction	• Did you get all the information you wanted using the chatbot? • Do you think the chatbot understood what you asked? • Overall, were you satisfied with the chatbot? • Do you think you would use this chatbot again?	[4]

The next section discussed the validation of the research instruments to determine the reliability and validity of the questionnaire prior to the release of the actual survey to the respondents.

IV. VALIDATION OF RESEARCH INSTRUMENTS

User experience testing was a process in which the interface and the chatbot features were verified by end users who execute specific tasks under practical environments. This test aimed to assess the user experience in terms of Usability, Usefulness and Satisfaction of the conversational agent and to ascertain if the application was functional. At the end of the testing, a survey was carried out via Google Form to gather user' response and satisfaction towards the usage of the conversational agent.

In addition to this, a pilot test has been carried out to ascertain the validity and reliability of the questionnaire survey whereby 60 respondents have been selected. According to [12], the number of respondents in the pilot test was determined by the total number of variables tested in the questionnaire. There were three items to be tested in the questionnaire which were Usability, Usefulness and Satisfaction upon the usage of the conversational agent. The following Table II summarized the descriptive analysis of the pilot test.

TABLE II. SUMMARY OF DESCRIPTIVE ANALYSIS

Descriptive Analysis	Respondent	Total
Age • 18 to 24 • 25 to 34	 59 (98.3%) 1 (1.7%)	60
Education Level • SPM	 60 (100%)	60
Occupation • Student	 60 (100%)	60
Used Conversational Agent before • Yes • No	 59 (98.3%) 1 (1.7%)	60

According to Table II, 60 respondents were students. 59 of the respondents were aged between 18 to 24 and 1 respondent was aged 25 to 34 years old. The highest educational level of the respondents was SPM and 59 respondents have experienced using conversational agent before. Furthermore, the reliability test and the validity test has been conducted and reported in Table III.

TABLE III. SUMMARY OF RELIABILITY TEST AND VALIDITY TEST

Measure	Cronbach's Alpha	KMO and Bartlett's Test
Usability	0.787	0.638**
Usefulness	0.780	0.559**
Satisfaction	0.793	0.663**
All items	0.895	0.720**

(** indicates the test was significant at 0.01 level)

The reliability test for each item has shown that the Cronbach's Alpha value for each item were more than 0.6. Furthermore, as for the KMO and Bartlett's test for validity test, all items in the questionnaire has achieved 0.720 and it was statistically significant at 0.01 level. Therefore, based on the reliability test and validity test, the questionnaire survey was suitable to be progressed to the actual survey.

V. SAMPLE SIZE POPULATION

The sample size population of respondents used in past studies pertaining to chatbot research was explored. In the study conducted by [13], a total of 169 users were participated in the study to investigate the impact of introducing language style to e-commerce chatbots to improve customer satisfaction, determined customer interest in the item and determined user interaction with the service provided by the conversational agent. Apart of this, a total of 105 participants engaged in a survey performed by [14] to test the chatbot customer service for the Venice Airport with the specially crafted modular system. A group of 101 undergraduates engaged in the study carried out by [15] to evaluate if the proposed novel paradigm enabled the users to nurture companion chatbot via developmental of artificial intelligence techniques.

Moreover, a total of 161 Korean students from major metropolitan universities in Korea engaged in research undertaken by [16] to indicate if the Chatbot e-service managed to provide interactive and engaging customer service encounters. Besides, in the test conducted by [17], 100 respondents were randomly chosen to signify the suggested

system based on certain abstract concepts, which could be applied to satisfy the necessary capabilities of the industry. In comparison, a total of 96 undergraduate computer science students engaged in research undertaken by [18] to recognise conversational agents for academically successful interactions, allowing learners to sustain effective peer dialogue in a range of learning environments.

The abovementioned evaluation on conversational agents indicated that the testing was carried out using non-probability sampling in which the approximate respondents ranging from 96 to 169 were used to carry out the testing. There was no clear generic measurement of the total number of respondents used. The sample size of this study was calculated based on the sample size formula [12], [19], refer to (1) to resolve this concern.

$$n = \frac{\left[\frac{z^2 * p(1-p)}{e^2}\right]}{\left[1 + \frac{z^2 * p(1-p)}{e^2 N}\right]} \quad (1)$$

Based on the sample size formula, n refers to the sample size, z denotes the z-score of confidence level, N denotes the population size, e denotes the margin of error, and p denotes the standard deviation. The confidence level is set at 95% with the z-score of 1.96. According to the estimation of the sample size from (1), a total of 300 users was selected to carry out the user experience testing.

VI. Testing and Analysis

The Demographic test results were reported in Table IV. The complete graph for the demographic was then explained further in this section. Based on the survey, 83% of the respondents aged between 18 to 24 are students with SPM as the highest education which constituent to 82.7%. Moreover, 97.7% of the respondents have used chatbot before. Next, Fig. 4 showed the summary of User Experience Testing captured for Usability, Usefulness and Satisfaction. There were total of four questions for each of the user experience parameters with the mixture of positive-typed questions and negative-typed questions to prevent random answer selection by users. The survey was capture via Likert-scale ranging from score-1 to score-5 to determine the average and standard deviation of the user experience.

In order to capture reliable data, the questions were formed with the mixture of positive-type-question and negative-type-question to prevent random answer selection by users. There were total of eight positive-type-question and four negative-type-question. The positive or negative indicators could be seen next to the question number in Fig. 4. The data was quantified via the Likert-scale, ranging from strongly disagree (score-1) to strongly agree (score-5). Table V showed the total average score for positive-type-question and negative-type-question.

The results in Table V and Fig. 5 depicted that question 1, question 3, question 5, question 7, question 9, question 10, question 11 and question 12 managed to achieve the total average score of 4.74 over 5. As these questions were positive-type-question and the average was achieved more than 4.7 which was towards the strongly agree score-5 in Likert-scale, this indicated users show positive experience towards the usage of the conversational agent. On the other hand, question 2, question 4, question 6 and question 8 were negative-type-question and each question managed to achieve the total average score of 1.24 over 5. Contrary to a positive-type-question, the lower number for negative-type-question in Likert-scale showed that users somehow deny the usage of the communication agent constitutes poor experience. Consequently, the findings of the User Experience Testing indicated that users were having positive experience of utilizing the conversational agent.

TABLE IV. Summary of Demographic Test Results

Question	Options	Percentage (%)	Number of respondents
What was your age?	Below 18	0.0%	0
	18 to 24	83.0%	249
	25 to 34	7.0%	21
	35 to 44	9.3%	28
	45 to 54	0.7%	2
	55 above	0.0%	0
What was your highest education?	Diploma	1.0%	3
	Degree	8.0%	24
	Master	6.7%	20
	PhD	1.0%	3
	STPM	0.0%	0
	SPM	82.7%	248
	Other	0.6%	2
What was your occupation?	Academician	8.0%	24
	Administrator	8.7%	26
	Programmer	0.0%	0
	Engineer	0.0%	0
	Designer	0.0%	0
	Salesperson	0.0%	0
	Businessman	0.0%	0
	Student	83.3%	250
	Other	0.0%	0
Have you ever used a chatbot before?	Yes	97.7%	293
	No	2.3%	7

TABLE V. Summary of user Experience Testing for Positive-Type-Question and Negative-Type-Question

Positive-type-question (+)		**Negative-type-question (-)**	
Question	**Average score**	**Question**	**Average score**
Q1	4.74	Q2	1.24
Q3	4.75	Q4	1.25
Q5	4.75	Q6	1.24
Q7	4.75	Q8	1.22
Q9	4.73		
Q10	4.71		
Q11	4.73		
Q12	4.73		
Total average	**4.74/5**	**Total average**	**1.24/5**

User experience parameters	No.	Questions	Type (Positive/ Negative)	Average	Standard deviation	Percentage (number of respondents)				
						1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree
Usability	Q1.	The chatbot's capabilities meet my requirements.	(+)	4.74	0.77	2.7% (8)	0.7% (2)	1.7% (5)	9.7% (29)	85.3% (256)
	Q2.	Using this chatbot was a frustrating experience.	(-)	1.24	0.80	89.7% (269)	4.0% (12)	2.0% (6)	1.7% (5)	2.7% (8)
	Q3.	This chatbot was easy to use.	(+)	4.75	0.74	2.0% (6)	1.3% (4)	1.7% (5)	9.7% (29)	85.3% (256)
	Q4.	I have to spend too much time correcting things with this chatbot.	(-)	1.25	0.80	88.7% (266)	5.0% (15)	2.0% (6)	1.7% (5)	2.7% (8)
Usefulness	Q5.	Because of this chatbot, I could quickly execute the task (retrieve answer).	(+)	4.75	0.74	2.0% (6)	1.3% (4)	1.7% (5)	9.7% (29)	85.3% (256)
	Q6.	This chatbot makes it hard to execute the task (retrieve answer).	(-)	1.24	0.80	89.7% (269)	4.0% (12)	2.0% (6)	1.7% (5)	2.7% (8)
	Q7.	Because of this chatbot, I could effectively execute the task (retrieve answer).	(+)	4.75	0.75	2.3% (7)	1.0% (3)	1.7% (5)	9.3% (28)	85.7% (257)
	Q8.	This chatbot was useless.	(-)	1.22	0.78	91% (273)	2.7% (8)	2.0% (6)	2.0% (6)	2.3% (7)
Satisfaction	Q9.	Did you get all the information you wanted using the chatbot?	(+)	4.73	0.76	2.3% (7)	1.0% (3)	1.7% (5)	11.3% (34)	83.7% (251)
	Q10.	Do you think the chatbot understood what you asked?	(+)	4.71	0.80	2.7% (8)	1.3% (4)	1.7% (5)	10.7% (32)	83.7% (251)
	Q11.	Overall, were you satisfied with the chatbot?	(+)	4.73	0.77	2.7% (8)	0.7% (2)	1.7% (5)	10.7% (32)	84.3% (253)
	Q12.	Do you think you would use this chatbot again?	(+)	4.73	0.83	3.3% (10)	0.7% (2)	2.0% (6)	8% (24)	86% (258)

Fig. 4. Summary of user Experience Testing.

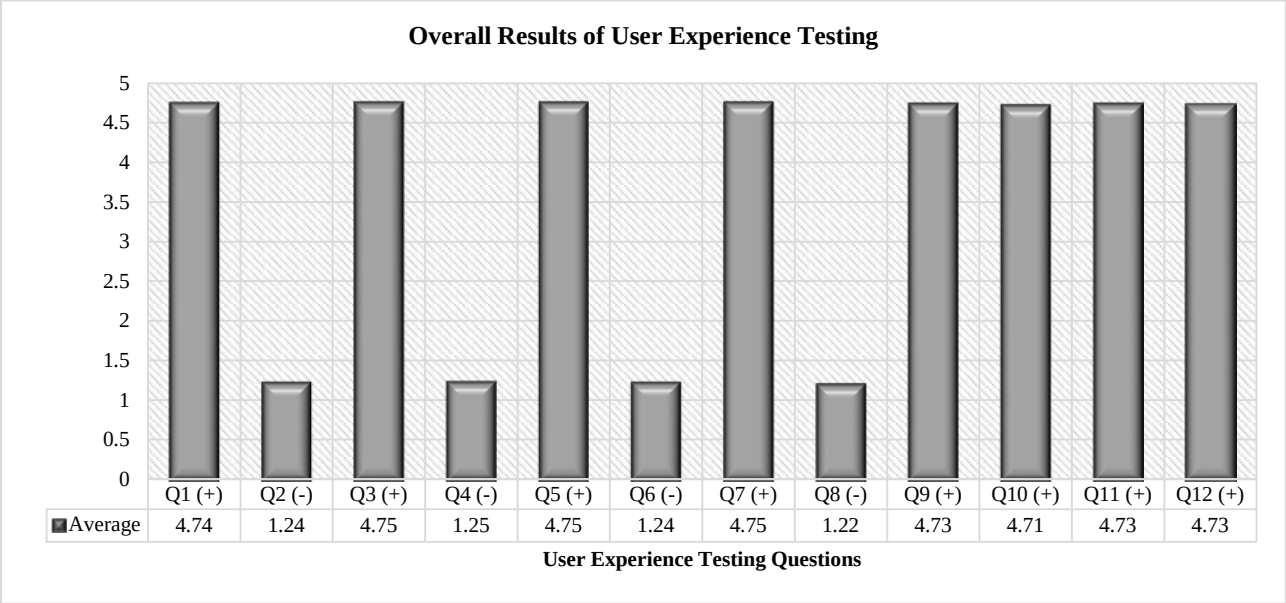

Fig. 5. Overall Results of user Experience Testing.

VII. Conclusion

Generally, the results from the testing indicate the success of the study. The findings from the positive-type-question managed to achieve the total average score of 4.74 over 5 against the widely accepted score-5 in Likert-scale, which suggested that users have good experience of using the conversational agent. The negative-type-question on the other hand managed to achieve the total average score of 1.24 over 5. In contrast with positive-type-question, the lower figure in Likert-scale for negative-type-question indicated that users somehow disagree that the usage of the conversation agent

constituent to bad experience. Apart from this, the test results from User Experience Testing indicated that 84.3% of the respondents were satisfied with the conversational agent, whereas 86% of respondents would use this conversational agent again. Thus, the results of the User Experience Testing demonstrated that users show positive experience towards the usage of the conversational agent. The analysis from User Experience Testing stipulated that most respondents were pleased with the conversational agent and would use it again. The respondents claimed that the chatbot was useful and they were able to retrieve answer quickly and effectively via this chatbot.

ACKNOWLEDGMENT

Special thanks to the Natural Language Computing Group, Microsoft Research Asia, Tunku Abdul Rahman University College (TAR UC), Universiti Teknikal Malaysia Melaka (UTeM) and MyBrain15 scholarship.

REFERENCES

[1] Duijst, D., 2020. What is the effect of personalization on the UX of Chatbots? [online] Available at: https:https://uxdesign.cc/what-is-the-effect-of-personalization-on-the-ux-of-chatbots 392bf34bba3b//www.researchgate.net/publication/318404775 [Accessed on 5 May 2020].

[2] Finstad, K., 2010. The Usability Metric for User Experience, Interacting with Computers, 22, pp. 323-327.

[3] Davis, F. D. (1985). A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results. Massachusetts Institute of Technology.

[4] Quarteroni, S., and Manandhar, S., 2008. Designing and Interactive Open-Domain Question Answering System, Natural Language Engineering, 1 (1), pp. 1-23.

[5] Orlando, T.M. and Sunindyo, W. D., 2017. Designing dashboard visualization for heterogeneous stakeholders (case study: ITB central library), 2017 International Conference on Data and Software Engineering (ICoDSE), Palembang, 1-2 November 2017, pp. 1-6.

[6] Rodriguez Gil, J. García-Zubia, P. Orduña, A. Villar-Martinez and D. López-De-Ipiña, 2018. New Approach for Conversational Agent Definition by Non-Programmers: A Visual Domain-Specific Language," IEEE Access, vol. 7, pp. 5262-5276.

[7] Chancusing, J.C., and Bayona-Ore, S., 2019. Information and Communication Technologies Acceptance Models in Universities. 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, pp. 62-66.

[8] Asastani, H.L., Harisno, V. H. Kusumawardhana and H. L. H. S. Warnars, 2018. Factors Affecting the Usage of Mobile Commerce using Technology Acceptance Model (TAM) and Unified Theory of Acceptance and Use of Technology (UTAUT), 2018 Indonesian Association for Pattern Recognition International Conference (INAPR), Jakarta, Indonesia, pp. 322-328.

[9] Setianto, F. and Suharjito, 2018, Analysis the Acceptance of Use for Document Management System Using Technology Acceptance Model, 2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, pp. 1-5.

[10] Harb, Y., and Alhayajneh, S., 2019. Intention to use BI tools: Integrating technology acceptance model (TAM) and personality trait model, 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, pp. 494-497.

[11] Natalia, S. Bianca and I. A. Pradipta, 2019. Analysis User Acceptance of Wonderful Indonesia Application Using Technology Acceptance Model (case study: Indonesian Ministry of Tourism), 2019 International Conference on Information Management and Technology (ICIMTech), Jakarta/Bali, Indonesia, pp. 234-238.

[12] Anderson, D.R., Sweeney, D.J., Williams, T.A., Camm, J.D., and Cochran, J.J., 2018. Statistics for Business and Economics, 13th Ed., Boston, USA: Cengage Learning.

[13] Elsholz, E., Chamberlain, J., and Kruschwitz, U., 2019. Exploring Language Style in Chatbots to Increase Perceived Product Value and User Engagement, Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR), Glasgow, UK, 10-14 Mar 2019, pp. 301-305.

[14] Carisi, M., Albarelli, A., and Luccio, F.L., 2019. Design and Implementation of an Airport Chatbot, EAI International Conference on Smart Objects and Technologies for Social Good (GoodTechs ’19), 25-27 September 2019, Valencia, Spain.

[15] Chen, G., 2018. Nurturing the Companion ChatBot. 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’18), 2-3 February 2018, New Orleans, LA, USA.

[16] Chung, M., Ko, E., Joung, H., and Kim, S.J., 2018. Chatbot e-service and Customer Satisfaction Regarding Luxury Brands, Journal of Business Research.

[17] Wei, C., Yu, Z., and Fong, S., 2018. How to Build a Chatbot Framework and its Capabilities, Proceedings of the 2018 10th International Conference on Machine Learning and Computing, Macau, China, February 2018.

[18] Tegos, S., Demetriadis, S., Papadopoulos, P.M., and Weinberger, A., 2016. Conversational Agents for Academically Productive Talk: A Comparison of Directed and Undirected Agent Interventions. International Journal of Computer-Supported Collaborative Learning, 11, pp. 417-440.

[19] Weiers, R.M., 2011. Introductory Business Statistics, 7th Ed., Boston, USA: Cengage Learning.