Position paper presented at CONVERSATIONS 2020 – the 4th International Workshop on Chatbot Research, an online virtual event hosted by the University of Amsterdam, the Netherlands, November 23-24.

Do Mistakes, Visual Elements, and Recommendation Explanation Affect Chatbot Usability? A Pilot Study.

Marianna A. de Sá Siqueira[1], Barbara Müller[1], and Tibor Bosse[1]

[1] Radboud Universiteit, Nijmegen, The Netherlands

m.desasiqueira@maw.ru.nl

Abstract. The exponential investment increase in chatbots demonstrates their current popularity and their strong potential for the future. Despite this high demand, chatbots still face the challenge of communicating in a natural, human-like manner, and errors are seen as a significant limitation of chatbots, which negatively impact users' experience. Based on the Technology Acceptance Model (TAM), the current pilot study investigates to what extent certain properties of chatbots (i.e., mistakes, visual elements, and recommendation explanation) have an impact on users' experience. Participants interacted with a chatbot twice, and subsequently completed questionnaires assessing the TAM concepts (i.e., Perceived Ease of Use, Perceived Usefulness and Attitude Towards Using), and two additional concepts: Enjoyment and Social Presence. In line with previous findings, preliminary results suggest that a chatbot that makes mistakes lead to a more negative user experience. Interestingly, correlation analyses reveal that Enjoyment and Social Presence relate positively to people's Attitude Towards Using a chatbot. Possible future directions are discussed.

Keywords: Chatbot Usability, User Experience, Technology Acceptance Model, Social Presence, Enjoyment

1. Chatbots and User Experience

Chatbots are computer-based programs that are able to simulate human-human conversation [1]. They are widely popular as a practical application. Nowadays, chatbots have several applications, such as customer service, virtual assistants, or even to chat about personal topics. Many companies already use chatbots as a tool for their customer contact, and if not yet, they are most probably planning to do so in the future [2]. Chatbots can be a useful tool because of their 24/7 availability and their reduction of expenses in companies. Especially chatbots used for customer contact are expected to deliver pleasant and valuable experiences to meet customer expectancies [3]. User Experience (UX) has become an important concept to consider in software development to help increase the usability and acceptance of chatbots. A good UX is a consequence of fulfilling the human needs for autonomy, competence, and stimulation through interacting with the chatbot [4].

To improve the user experience, in the current study we introduce three factors that were manipulated in the context of a chatbot-based recommendation system, namely: *chatbot mistakes*, *visual elements*, and *recommendation explanation*. Focusing on these three factors, the aim of the study was to analyse how these different features impact a chatbot's usability. In particular, we investigate to what extent the mistakes decrease and the visual elements and recommendation explanation improve the usability.

Although chatbots have been in use for a long time, and their quality is improving over the years, their performance is still not perfect. In 2020, chatbots still need adjustments and improvements to become the preferred tool for most users [5]. Errors are still considered to be a significant limitation of chatbots. For example, they may not recognize the user's intention or provide inappropriate answers during a conversation, which negatively impacts users' experience and leads to disappointment [6]. The ability to provide correct answers is thus essential for chatbots, in particular in the context of recommendation systems [7]. Hence, analyzing the effects of such errors would be helpful for developers to get people to use their chatbots frequently [8]. By clarifying whether certain types of mistakes (e.g., too long pauses, wrong recommendation, disability to answer a question) have different effects on chatbot use, it would be possible to prioritize future chatbot development.

Visual elements are sources of information in the form of visual representations, such as photographs, videos, and animations [9]. The power of using visual information together with textual information is to physically represent objects or actions and visualize a scenario, tell a story, or highlight specific information [8,10]. When considering chats, emojis and GIFs are two available visual elements that carry emotional information, and emoji keyboards make the use of emojis even easier for smartphone users [11,12].

With the advent of recent machine learning algorithms, the ability to explain individual decisions is currently one of the 'hot topics' in Artificial Intelligence. According to Teach et al. [13], the ability to explain advice is the most critical quality of a computer-based decision support system. Moreover, the explanation component in expert systems is claimed to improve user satisfaction, user acceptance, and perceived quality of the system [14]. Research has shown that a chatbot is perceived as more human-like when it explains its answer [15].

The current paper aims to answer the question: Do mistakes, visual elements, and explanations affect chatbot usability?

1.1. Technology Acceptance Model

In this research, we make use of the Technology Acceptance Model (TAM) as a theoretical basis. This model is mainly used to investigate the effect that certain features of a system have on users' intentions, beliefs, and attitudes. The model suggests that two factors influence a user's Attitude Towards Using technology: Perceived Ease of Use and Perceived Usefulness [16].

Perceived Ease of Use is "the degree to which a person believes that using a particular system would be free of effort" [17]. Perceived Usefulness is "the degree to which a person believes that using a particular system would enhance his or her job performance" [17]. To optimally benefit from the system, its use should be comfortable and effortless. Attitude Towards Using the chatbot is a user's general impression of the system. A system with high Perceived Ease of Use and Perceived Usefulness will lead to a more positive Attitude Towards Using the chatbot the system.

Next to the TAM, two additional concepts were included in our work: Social Presence and Enjoyment. The first is an important concept when it comes to trust towards technologies. Blau [18] discovered that trust is built within a social environment, and therefore a feeling of social presence is necessary to develop trust. When communicating, emotional information shown via facial expression, posture, and non-verbal cues contributes to the social presence [19], and enjoyable conversations increase the feeling of others being present [20]. Logically, social presence is highest in face-to-face communication [21]. Enjoyment is a concept that measures the perceived enjoyment of a system apart from any performance consequences that may be anticipated [22]. An online system that is enjoyable to use is likely to be used more frequently and has a higher probability of being used again [23,24]. Davis et al. [25] mentioned Enjoyment as an intrinsic motivator to use technology, while the TAM concepts of Perceived Ease of Use and Perceived Usefulness are extrinsic motivators. Enjoyment influences a user's attitude towards using a system and is thus an essential concept for measurement.

Combining the TAM, Social Presence, and Enjoyment, our conceptual model is depicted in Figure 1. X represents a specific feature of the chatbot (i.e., chatbot mistakes (1), visual elements (2), recommendation explanation (3)), and H_X represents the hypothesis that feature X has an effect on user experience, measured in terms of Perceived Ease of Use, Perceived Usefulness, Social Presence, and Enjoyment, respectively. Furthermore, these four aspects of user experience are hypothesized to have an effect on the Attitude towards using the chatbot.

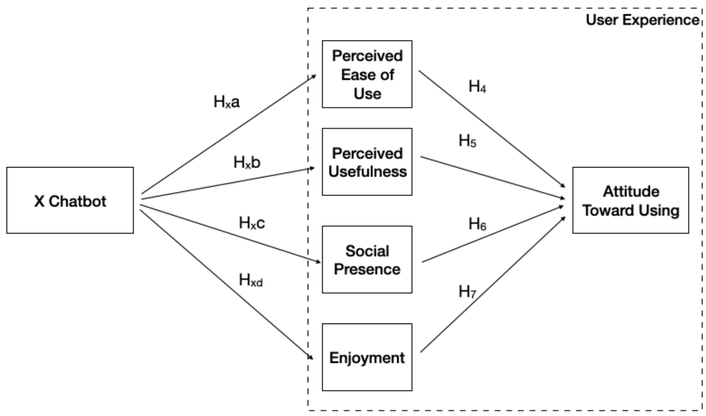

Fig. 1. Research Model based on Technology Acceptance Model [17].

The hypotheses are described below:

H_1a: A chatbot that does not make mistakes is perceived as easier to use than a chatbot that makes mistakes.

H_1b: A chatbot that does not make mistakes is perceived as more useful than a chatbot that makes mistakes.

$H_{1}c$: A chatbot that does not make mistakes is perceived as having a higher social presence than a chatbot that makes mistakes.

$H_{1}d$: A chatbot that does not make mistakes is perceived as more enjoyable than a chatbot that makes mistakes.

H_{2} and H_{3} are comparable, but they consider Visual Elements and Explanation as context variables which are expected to have a positive effect.

H_{4}: The higher the Perceived Ease of Use of a chatbot, the more positive people's attitude toward using it.

H_{5}: The higher the Perceived Usefulness of a chatbot, the more positive people's attitude toward using it.

H_{6}: The higher the Perceived Social Presence of a chatbot, the more positive people's attitude toward using it.

H_{7}: The higher the Perceived Enjoyment of a chatbot, the more positive people's attitude toward using it.

2. Research Methodology

The chatbot used for this pilot study is a movie recommendation system, which gave recommendations based on a score which was acquired during the dialogue with the user's preferences. Users had some interactions with the chatbot and were asked for movie preferences until the chatbot had enough information to give a recommendation. The dialog was developed using Dialogflow. The user interface was built with the use of the JavaScript library React. This library is beneficial when working with an interactive user-interface. Python was used to implement the algorithms for finding the right recommendation based on the user's preferences. We used The Movie Database (TMDb) API to search for movies that matched the preferences. Users accessed the chatbot through a URL address.

Default chatbot. The default chatbot was developed without any intentional mistakes, no visual information, no memory of the previous conversation, and no recommendation explanations. The default conversation has some human-like interaction and some interaction contact, and it presents itself as "Chad." All following chatbots were developed based on this default conversation, with additional adjustments on the dialog.

Chatbot mistakes. To implement intentional mistakes, adjustments in the back end were made. Firstly, the recommendation algorithm in python no longer receives movies that satisfy the user's genre preference. Instead, it receives only movies that do not satisfy this preference. Subsequently, the scores are computed in the same way as they are for the original chatbot. However, this time the movie with the worst score is returned to the user. This is possible by sorting the movies in ascending order. Note that we include the question for the movie's genre, to demonstrate that the answer is wrong.

Visual elements. For this feature, emojis were added to the already existing responses to support a message. Some affirmative responses like 'okay, good to know' were replaced by GIFs (retrieved from www.giphy.com) that showed that the message was understood (e.g., seeing someone nodding in the GIF). Some GIFs were added to the responses to be funny and therefore supposedly making the conversation more enjoyable. The use of emojis and GIFs was restricted as overusing them can make people get bored and give them the feeling that their chat partner tries to hide their real intentions [26].

Recommendation explanation. The only difference with the default chatbot is that when the user receives the movie recommendation, in addition to the movie title, the user receives a brief explanation about which information was used by the recommendation engine to give him/her that response.

The user experience in all conditions is measured by the combination of TAM, Social Presence, and Enjoyment elements, as shown in figure 1.

2.1. Study Design & Procedure

Data collection was during 15 days in May of 2020, and it was entirely online. We had 54 participants in total (59.9% female); 16.8% of them were between 18-20 years, 62% between 21-25 years, and 21.2% above 25 years. 60% of the participants had a Bachelor degree, 29% had a Master degree, and 11% completed secondary school. They were randomly assigned to the 4 between-subjects conditions. Participants either interacted with a default chatbot, a chatbot making mistakes (i.e., giving the wrong recommendation), a chatbot using visuals and gifs, or a chatbot explaining its decisions. All participants were recruited through social media networks, that is, through Facebook, using the social network of students who helped with data collection, and by posting the experiment on different students' group pages. They received the instruction to interact with the chatbot, receive one movie recommendation, and extract at least two pieces of information about the movie that was recommended. They could ask about the movie score, cast, synopsis, genre, and duration. They were instructed to use the chatbot twice, interacting with it the same way following the conversation as it was presented, with the second time being (preferably) one day after the first time to cover for the novelty effect (i.e., a positive effect on performance due to the introduction of new technology [27]). Each participant interacted with the same chatbot presenting the same conversation flow. It was not explicitly mentioned to ask for a different movie recommendation on the second interaction. Subsequently, to measure the user experience after each interaction, participants completed questionnaires on Attitude Towards Using [28] (5 items), Perceived Usefulness [28] (5 items) and Perceived Ease of Use [28] (3 items), and the additional concepts of Social Presence [29] (4 items) and Enjoyment [30] (3 items). This pilot had a mixed design, with a chatbot feature (default vs. mistake vs. visual vs. explanation) as a between-subjects factor, and time (T1 vs. T2) as a within-subjects factor (figure 2 provides an overview of the experiment timeline). Figure 3 shows the layout of each chatbot displayed on a mobile screen.

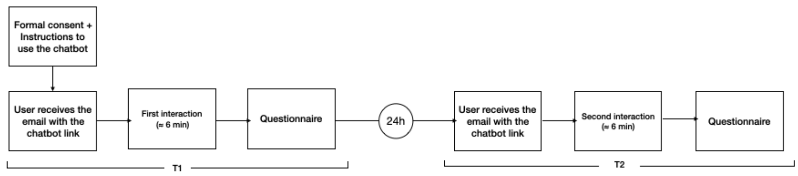

Fig. 2. Experiment timeline.

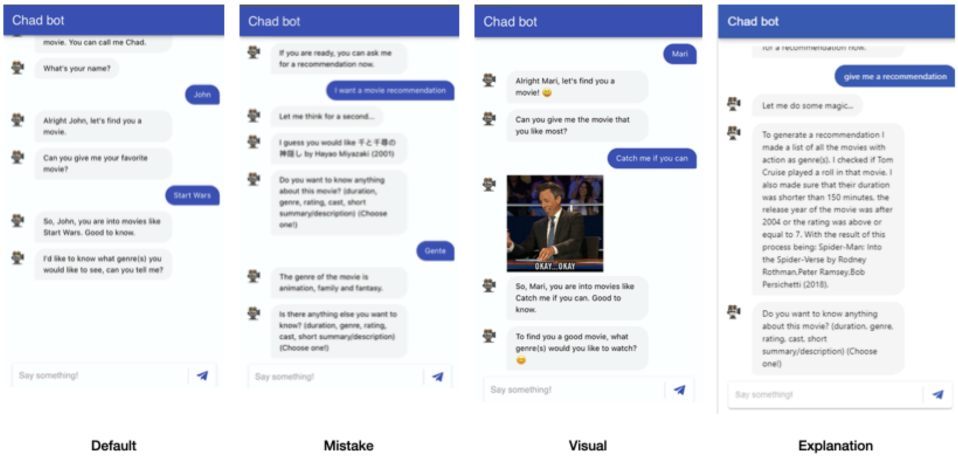

Fig. 3. Chatbot screenshot examples - from left to right - of the dialog for the four conditions default, mistakes, visual information, and recommendation explanation, respectively.

3. Results

Table 1 shows the means of the measured variables for each chatbot condition.

Table 1. Mean and standard deviation T_2.

	Default (N=16)	**Mistakes (N=16)**	**Visual (N=13)**	**Explanation (N=9)**
Perceived Ease of Use	4.27 (±0.43)	4.44 (±0.48)	4.00 (±0.56)	4.19 (±0.24)
Perceived Usefulness	3.50 (±0.78)	2.74 (±1.32)	3.28 (±0.70)	3.49 (±0.79)
Social Presence	3.47 (±0.75)	3.20 (±0.86)	3.06 (±0.74)	3.44 (±0.39)
Enjoyment	3.25 (±0.82)	2.94 (±1.12)	2.95 (±0.77)	2.19 (±0.69)
Attitude Towards Using	3.59 (±0.86)	3.31 (±1.01)	3.46 (±0.72)	3.58 (±0.71)

Due to difficulties to link T_1 with T_2 data[1], we decided to only analyse the data of T_2. We expected that these data were less confounded by novelty of interacting with a chatbot. We used one-way ANOVAs on the T_2 data, which revealed no significant effects for most of the conditions (all *p's* > .1), and thus, hypotheses H_1a, H_1c, H_1d, H_2, and H_3 could be rejected. However, a marginally significant effect was found when comparing the Default chatbot with the chatbot making mistakes on Perceived Usefulness as a dependent variable [$F(1,30) = 3.96, p = .056$]. The default chatbot was considered more useful (M_{ud}= 3.5) than the chatbot making mistakes (M_{uw}= 2.74). So, H1b can be conditionally accepted.

Pearson correlations were calculated to test H4, H5, H6, and H7. We found a moderate positive correlation between Perceived Ease of Use and Attitude Towards Using the chatbot [$r(134)=.37, p<.01$]. There was a high positive correlation between Perceived Usefulness and Attitude Towards Using the chatbot [$r(134)=.80, p<.01$]. There was also a high positive correlation between Social Presence and Attitude Towards Using the chatbot [$r(134)=.56, p<.01$]. Lastly, there was a high positive correlation between Enjoyment and Attitude Towards Using the chatbot [$r(134)=.57, p<.01$]. Therefore, H_4, H_5, H_6, and H_7 were accepted.

4. Discussion

In this project, we aimed to evaluate the user experience of chatbots by manipulating three factors in a movie recommendation system: mistakes, visual elements, and recommendation explanation. Results suggest that while H_1a, H_1c, H_1d, H_2, and H_3 were rejected, H_1b, H_4, H_5, H_6, and H_7 were accepted. In other words, the Perceived Usefulness scores for the chatbot with mistakes were significantly lower than for the default chatbot. Furthermore, strong positive relations between Perceived Ease of Use, Perceived Usefulness, Social Presence, and Attitude Towards Using the chatbot were found. Given the low number of participants, and the fact that ANOVAs need more participants to be able to generalize findings, the current results have to be interpreted with caution and more data is necessary to draw clear conclusions. According to comparable studies [5], the current study should have at least 240 people distributed across four conditions. With a larger sample, it could therefore be possible that we are able to confirm our hypotheses for mistakes and explanation, which currently only show a trend in the expected direction.

Does the mistake level affect chatbot usability? Research has shown that computer mistakes (represented as a wrong recommendation in this study) are frustrating to users [29]. Despite having a lower Perceived Usefulness score for the mistake chatbot, we had a slightly better evaluation for the mistake chatbot on Perceived Ease of Use. This can possibly be explained by the fact that the mistake was only in one part of the conversation, and thus can be "forgiven" by the user. Future research should investigate

[1] Participants did not have a user name or number to be able to follow them over the two timepoints. We tried to do so using IP address, but this procedure did not guarantee that each address belonged to the same user.

other types of mistakes (e.g., misunderstanding of the question, not responding) to be able to clarify how chatbot mistakes can further affect the user experience. For example, if the mistake involved presenting results that are systematically wrong, such as being in the wrong order or category, the output would appear structured to an experienced user. While systematic errors are possible to repair, equally disturbing are errors that do not make sense and are impossible to interpret by the user. As the interaction mode with chatbots leans heavily on the human practice of natural conversations, one would expect that this applies particularly to chatbots: usually, users have strong expectations about how the interaction should take place. Thus, future chatbot development could be prioritized by clarifying whether certain types of mistakes (e.g., too long pauses, wrong recommendation, disability to answer a question) have different effects on chatbot use.

Do visual elements and explanations affect chatbot usability? According to the literature, visual emotional information strengthens a conversation because of its ability to express personality and convey humor [17,20,30]. Looking at the means in our study, it seems this information had a worse evaluation in most cases compared with the default chatbot. One of the reasons can be the overuse of this feature in a short dialog like in the present study. It would be interesting to see in a future study whether certain thresholds can be identified to avoid an overuse of visual elements. The explanation of recommendations is a relevant topic in the context of acceptance of technology [31]. Looking at the means, we can see a better evaluation when the chatbot gives an explanation. It might be that in other domains, such as medical diagnosis, explanations have a higher relevance and lead to stronger effects. To further explore the effect of explanations on chatbot usability, it would therefore be advisable to use another domain where more information about the decision process is appreciated.

An interesting finding of this study is the high correlation of Enjoyment with the Attitude Towards Using the chatbot. As reported by Davis et al. [25], Enjoyment is an intrinsic motivator for system use, and this is supported by our pilot. Moreover, as we expected, the other three concepts (Perceived Ease of Use, Perceived Usefulness, and Social Presence) positively correlate with the Attitude Towards Using the chatbot. Thus, generally it can be concluded that factors that increase Enjoyment, Perceived Ease of Use, Perceived Usefulness, and Social Presence, increase Attitudes towards using the Chatbot. Looking at the complete TAM, this could then have a positive influence on Intention to use the chatbot, and eventually chatbot use.

In this pilot, we only consider the user experience using the TAM, Social Presence, and Enjoyment, but other models would also be interesting to measure the usability, for example the UTAUT2 model [34]. Important to note, the relationship among the three factors (mistakes, visual information, and explanation) was not tested in the current pilot. A larger sample would make it possible to investigate theses possible interactions to verify whether and how these factors interact.

The current pilot is part of a project that analyses the communication features of a chatbot that make mistakes. The next step is to improve the conversation and test some of these features against two different 'mistake levels' of the chatbot (low and high). In addition, we plan to increase the sample size to be able to draw firm conclusions. Based on our experiences in this pilot, we will also add qualitative questions to the questionnaire to be able to better understand what people expect from a chatbot conversation before and after the interaction. Our pilot provides insights into the role of communication features in human-chatbot conversation and hopefully inspires future studies to investigate how to further improve chatbot interactions.

5. Acknowledgments

We would like to thank Hortense Verhoeven, Leon Driessen, and Thijs Luttikholt for their help in data collection for this pilot study.

References

1. Cambridge Dictionary: Chatbot. In Cambridge Dictionary. Retrieved November 2, 2020, from https://dictionary.cambridge.org/dictionary/english/chatbot (n.d.)
2. Wrangler, T.: Chatbot adoption report, https://es.scribd.com/document/373850606/Chatbot-Summary-2018- Report-TCW-PCAuthoring-Solutions-Draft-March-7-2018, last accessed 2020/04/20.
3. Jassova, B.:Recommendation Chatbot for eCommerc, https://landbot.io/blog/product-recommendation-chatbot-for-ecommerce/, last accessed 2020/04/20.
4. Hassenzahl, M.: User experience (UX): towards an experiential perspective on product quality. In: 20th International Conference of the Association Francophone d'Interaction Homme-Machine, pp. 11–15. Metz, France (2008).
5. Araujo, T.: Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perception. Computers in Human Behavior 85, pp. 183-189 (2018).
6. Luger, E., Sellen, A.: Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 5286-5297 (2016). ACM. http://dx.doi.org/10.1145/2858036.2858288
7. Shani, G., Gunawardana, A.: Evaluating recommendation systems. In: F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), Recommender systems handbook, pp. 257–297, Boston, MA: Springer US (2011). doi: 10.1007/978-0-387-85820-3 8
8. Preusse, K. C.: How people interpret and react to everyday automation issues (Unpublished doctoral dissertation). Georgia Institute of Technology (2016).
9. Spacey, J.: 11 Types of Visual Information, https://simplicable.com/new/visual-information, lasted accessed 2020/04/20.
10. Dingwall, J., Labrie, C., McLennon, T., Underwood, L. Professional Communication. Ontario: eCampus Ontario, https://ecampusontario.pressbooks.pub/profcommsontario/chapter/ using-visuals/#chapter-1111-section-8, last accessed 2020/02/16.
11. Betters, E.: Google adds SMS to Hangouts Android App, Emoji to KitKat keyboard, https://www.pocket-lint.com/apps/news/google/124960-google-adds-sms-to-hangouts-android- app-emoji-to-kitkat-keyboard, lasted accessed 2020/01/30.
12. Dimson, T.: Emojineering part 1: machine learning for emoji trends, https://instagram-engineering.com/emojineering-part-1-machine-learning-for-emoji-trendsmachine- learning-for-emoji-trends-7f5f9cb979ad, lasted accessed 2020/05/12.
13. Teach, R. L., Shortliffe, E. H.: An analysis of physician attitudes regarding computer-based clinical consultation systems. Computers and biomedical research, an international journal, 14(6), 542–558 (1981).
14. Darlington, K.: Aspects of Intelligent Systems Explanation. Universal Journal of Control and Automation 1(2), pp. 40-51, (2013) http://www.hrpub.org DOI: 10.13189/ujca.2013.010204.
15. Følstad, A., Nordheim, C. B., Bjørkli, C. A.: What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In: International conference on internet science, pp. 194–208 (2018).
16. Legris, P., Ingham, J., Collerette, P.: Why people use information technology? A critical review of the technology acceptance model. Information & Management, 40(3), 191-204 (2003).
17. Davis, F.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319-240 (1989).

18.Blau, P.: Exchange and Power in Social Life. New York: Wiley (1964).
19.Short, J., Williams, E., Christie, B.: The Social Psychology of Telecommunications. London: Wiley (1976).
20.Hsieh, S., Tseng, T.: Playfullness in mobile instant messaging: Examining the influence of emoticons and text messaging on social interaction. Computers in Human Behavior, 405-414 (2017).
21.Gefen, D., Straub, D.: Managing user trust in B2C e-services. e-Service Journal, 2(2), pp 7-24 (2003).
22.Hassanein, K., Head, M.: Manipulating perceived social presence through the web interface and its impact on attitude towards online shopping. International Journal of Human-Computer Studies, 65(8), 689-708 (2007).
23.Teo, T., Kim, V., Lai, R.: Intrinsic and extrinsic motivation in Internet usage. Omega, pp. 25-37 (1999).
24.Koufaris, M., Kambil, A., Labarbera, P.: Consumer behavior in Web-based commerce: An emperical study. The International Journal of Electronic Commerce, pp. 115-138 (2002).
25.Davis, F., Bagozzi, R., Warshaw, P.: Extrinsic and Intrinsic Motivation to Use Computers in the Workplace. Journal of Applied Social Psichology, 22,14, pp. 1111-1132 (1992).
26.Yoo, J.: To smile or not to smile:): Defining the effects of emoticons on relational outcomes. In Annual meeting of International Communication Association Conference (2007).
27.Koch, M., von Luck, K., Schwarzer, J., Draheim, S.: The novelty effect in large display deployments - Experiences and lessons-learned for evaluating prototypes. Proceedings of the 16th European Conference on Computer-Supported Cooperative Work-Exploratory Papers (2018).
28.Armentano, M., Christensen, I., & Schiaffino, S.: Applying the Technology Acceptance Model to Evaluation of Recommender Systems. Polibits, 51, 73-79 (2005).
29.Lowenthal, P. R.: Social Presence: What is it? How do we measure it? (Unpublished doctoral dissertation) University of Colorado Denver (2012).
30.Chung, M., Ko, E., Joung, H., & Kim, S.: Chatbot e-service and customer satisfaction regarding luxury brand. Journal Of Business Research (2018).
31.Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J., Shneiderman, B.: Determining causes and severity of end-user frustration. International Journal of Human–Computer Interaction, 17 (3), 333-356 (2004). doi: 10.1207/s15327590ijhc1703\ 3
32.Miltner, K. M., Highfield, T.: Never gonna GIF you up: Analyzing the cultural significance of the animated GIF. Social Media + Society, 3(3) (2017).
33.Ye, L. R., Paul E. J.: The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice. MIS Quarterly 19(2), pp. 157-72 (1995). doi:10.2307/249686.
34.Venkatesh, V., J. Thong, & X. Xu.: Consumer acceptance and user of information technology: Extending the unified theory of acceptance and use of technology. MIS Quarterly, 36 (1), 157-178 (2012). DOI: https://doi.org/10.2307/41410412