Continuous Improvement Study of Chatbot Technologies using a Human Factors Methodology

Abstract ID: 2868

Abstract

The emergence and pervasiveness of the Internet provides opportunities for new types of communications between customers and service providers. One such technology is a chatbot: a computer program that simulates a human conversation enabled by the Internet. Chatbots are currently used for a variety of reasons, from daily weather reports to ordering pizza! In this paper we present the investigation of chatbot technologies for an industry partner to better their internal communication process between field technicians and engineers. Implementation of this technology will automate the technician-to-engineer communication process and thus will result in a much more efficient system. Our team followed a human factors engineering methodology where we compared different chatbot platforms (IBM, Pandora, Self-development Kit) via usability testing. User feedback was derived systematically from a wide range of users and usability of several platforms was tested using a cognitive walkthrough and the data was quantified using a System Usability Scale (SUS). The majority of the participants in this research study (eight out of 10) preferred IBM's Watson. This platform received an average SUS score rating of 81.9 out of 100.

1.0 Background

Many companies are integrating chatbots into their websites to provide better user experience. What this means is that millions of people can communicate with your brand without a human being on the other end, overall saving the company time and money. So, if a company decides to implement a chatbot on their website, how do they know what kind of chatbot would be the most useful to them?

Chatbots are created using artificial intelligence (AI), which is the algorithm behind its ability to mimic a human conversation. There are two forms of AI that are currently being used along with this technology: machine learning (ML) and deep learning (DL). Per an article written by Michael Copeland on the NVIDIA website, ML uses rules to program a chatbot and is limited in terms of variability. The technology must be continuously monitored because it is only as smart as it is programmed to be. The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension -- as is the case in data mining applications -- machine learning uses that information to detect patterns in data and adjust program actions accordingly. DL is the newer form of ML that is becoming increasingly popular, but it is much more difficult to implement. In DL, a bot uses neural networks that were developed based on the understanding of how our brains work. But, unlike a biological brain where any neuron can connect to any other neuron within a certain physical distance, these artificial neural networks have discrete layers, connections, and directions of data propagation. [1]

No matter what form of learning the chatbot uses, the primary function of a chatbot is to serve the user. The way in which they are utilized can vary between facilitating human communication to completely replacing the need for humans to communicate with each other at all. The goal in implementing chatbot technology is to first and foremost save an organization time and money. The most challenging task for the organization will be to decide what type of chatbot fits their needs the best, whether it be the use of machine learning or the incorporation of deep learning as well. This will ultimately depend on what the chatbot will be used for and how much interaction the company wants the chatbot to have with the user.

2.0 Problem Identification

A major telecommunication company in the United States desired a more efficient method of communication between their field technicians and engineers. They believed that this would best be solved with the use of a chatbot. Our project was to give this company a recommendation of a chatbot technology and design for their internal communication system. To evaluate the usability of this technology, we took a human factors approach to this problem and designed an original research study to determine which chatbot platform users preferred the most for everyday use.

We designed a research study that would include user interviews, usability testing sessions with think-aloud protocols, and surveys/questionnaires. To help us achieve our objective, we compared IBM's Watson, Pandora's Pandorabot, and our very own Verbot from an available self-development kit. We programmed these underlying chatbot platforms to respond to yes or no questions with magic eight-ball-style answers. This would enable the participants in this study to focus on the usability and human factors aspects associated with the platform, instead of evaluating each chatbot on the accuracy of its answers. Based on our results, we were able to recommend a user-preferred platform. In this paper, we report our methodology and results and provide recommendations on how a company should adopt a chatbot to aid potential future adopters.

3.0 Methods

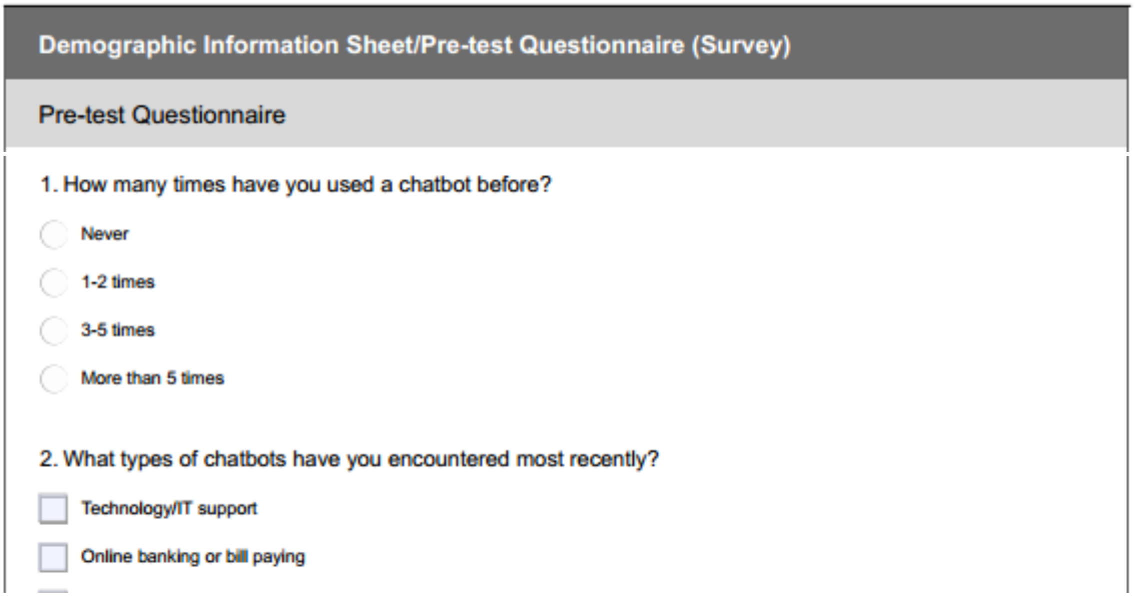
A usability test was conducted to evaluate the usability of three chatbots: IBM's Watson, Pandora's Pandorabot, and Verbot; a chatbot we developed from an available self-development kit. Each member of our team completed Texas A&M University's Collaborative Institutional Training Initiative (CITI) training and certification course and the project received IRB approval to enable us to conduct this study on human subjects.

3.1 Participants

Research shows that usability testing with just eight participants unveils about 80% of major usability issues [2]. A total of ten participants, with a mean age of 25.5 and standard deviation of 5.74, agreed to take part in this original research study. In an effort to diversify the range of user experience and perspective, we chose to include varying numbers of undergraduate students, graduate/P.H. D. students, and staff from the Texas A&M University community.

3.2 Procedure

After we gathered demographics information from the participants, we asked that each of them sign a consent form stating that they allowed us permission to video and/or audio record their chatbot session. Then, we began by asking each participant to answer a series of pre-test questions purposely designed to help us further explore each user's preconceived ideas regarding topics such as what kind of color scheme the user found more visually appealing (Would you prefer (1) a color scheme typically thought to invoke relaxation/stress relief, or (2) a more vibrant color scheme?). Along the same lines, we inquired as to whether or not each user would prefer to interact with an animated representation of the chat service, or avatar. The pre-test questionnaire is shown below in **Figure 1**.



Demographic Information Sheet/Pre-test Questionnaire (Survey)

Pre-test Questionnaire

1. How many times have you used a chatbot before?

○ Never

○ 1-2 times

○ 3-5 times

○ More than 5 times

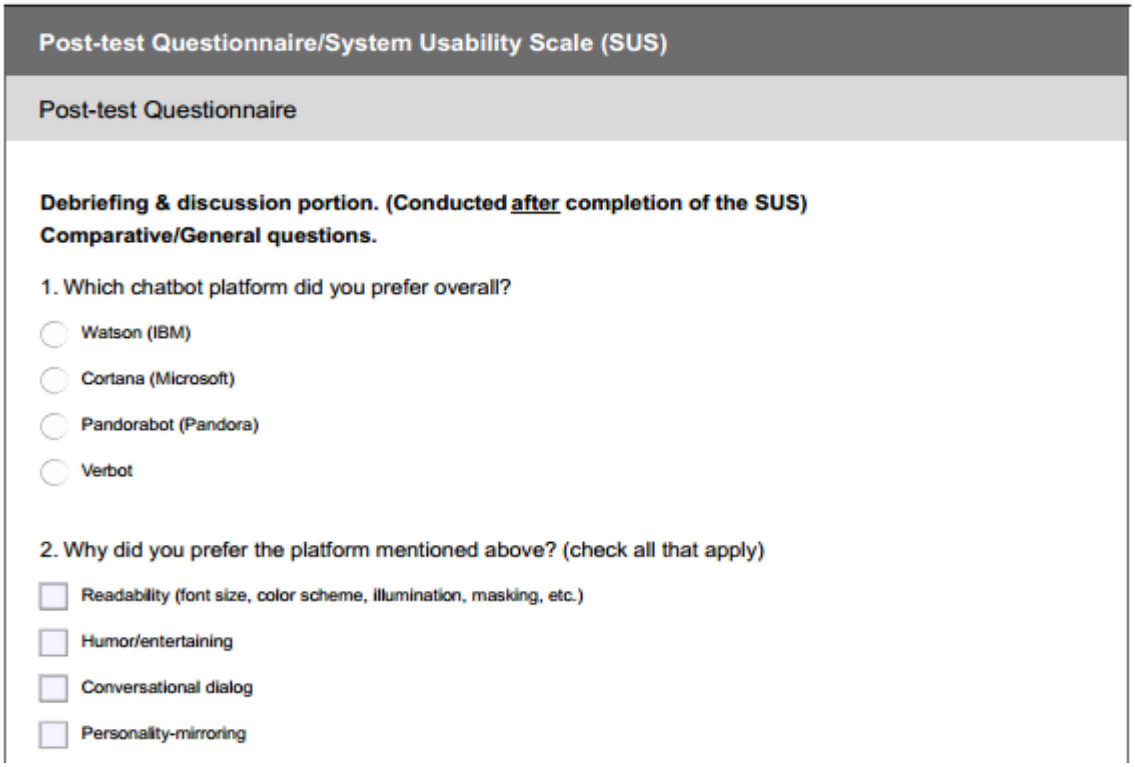
2. What types of chatbots have you encountered most recently?

□ Technology/IT support

□ Online banking or bill paying

Figure 1: Pre-test Questionnaire

Next, we developed our post-test questionnaire and a modified System Usability Scale (SUS) to assess usability of the different chatbot platforms. In particular, we asked participants to verbalize their responses to the questions and justify their answers. The post-test questionnaire is shown below in **Figure 2**.



Post-test Questionnaire/System Usability Scale (SUS)

Post-test Questionnaire

Debriefing & discussion portion. (Conducted after completion of the SUS)
Comparative/General questions.

1. Which chatbot platform did you prefer overall?

○ Watson (IBM)

○ Cortana (Microsoft)

○ Pandorabot (Pandora)

○ Verbot

2. Why did you prefer the platform mentioned above? (check all that apply)

□ Readability (font size, color scheme, illumination, masking, etc.)

□ Humor/entertaining

□ Conversational dialog

□ Personality-mirroring

Figure 2: Post-test Questionnaire

We calculated the SUS score of each of the participant's preferred chatbot platform so that we could quantify this data and make side-by-side comparisons. The SUS questions were asked with a rating of 1 to 5 with 1 representing "strongly disagree" and 5 representing "strongly agree". The ten SUS questions are shown below:

1. I think that I would like to use this Chatbot frequently.
2. I found the platform unnecessarily complex.
3. I thought the platform was easy to use.
4. I think that I would need the support of a technical person to be able to use this Chatbot.
5. I found the various functions in this platform were well integrated.
6. I thought there was too much inconsistency in this platform.
7. I would imagine that most people would learn to use this Chatbot very quickly.
8. I found the platform very cumbersome to use.
9. I felt very confident using the Chatbot.
10. I needed to learn a lot of things before I could get going with this Chatbot.

4.0 Results and Discussion

3.1 Pre-test Questionnaire

We discovered before we had even begun conducting these sessions that most people, if given the opportunity, would be more likely to call a company's service representative on the phone rather than attempting to use an automated chat service to find answers to their questions. We took a poll to determine which aspects of this technology were essential to how user-friendly this online service was perceived. Most users agreed that conversational dialog, the level of humor displayed and how entertaining a chatbot was, and the ability of a chatbot to adapt to a user's speech patterns, vocabulary, etc. (typically referred to as personality-mirroring) were essential factors in determining whether or not a user thought that they could benefit from using this technology. The pre-test questionnaire also revealed that while most people would describe their previous chatbot encounters to be positive overall, few had given much consideration as to how they felt this automated service had performed across different spheres of functionality. Therefore, this was the first opportunity many of these participants had to develop specific preferences regarding this technology. A lack of previous knowledge proved to be a beneficial factor in this study, as it contributed to the uniformity and the validity of our results.

3.2 Interaction With Chatbots

Eight out of ten participants chose Watson as their preferred platform while two users chose Pandorabot and not one of the ten users opted in favor of Verbot. Most participants found IBM's Watson to rank the highest in terms of readability and visual appeal, while Watson and Pandorabot tied in the category to determine which of the three was the most entertaining. Watson and Pandorabot were both praised for their sharp color contrast. This factor was considered to be more ergonomically beneficial, and therefore better in terms of readability for the user.

Most users disagreed in the evaluation of Pandorabot and Verbot's avatars. Some users claimed that on a professional level, if this technology were to be used in everyday communications between technicians and engineers, a simple chat window would suffice. Others, however, said that the use of an avatar made the technology more user-friendly by allowing the user to have a more entertaining interaction with the chat service. It was clear, given their responses, those users overwhelmingly preferred to interact with Pandorabot's avatar, rather than interacting with Verbot. Six of the ten participants stated that Verbot's monotone voice and its lag in response time gave users the impression that the platform was too robotic when responding to questions. These participants claimed that they did not enjoy the feeling of talking to a computer. On the other hand, a majority of particpants liked Pandorabot because they claimed it had a high level of human-like functionality, due to the fact that its voice sounded less robotic, in addition to providing more humorous responses. In other words, Pandorabot was better suited to entertain the user.

IBM's Watson chatbot platform had no avatar, had contrasting colors, and used a machine learning algorithm as well as a cloud-based database. One participant in our study spoke in depth about how Watson's sense of perceived intelligence was felt in the accuracy of the answer and that it made the user more confident and comfortable using this technology. This factor also contributed to the platform being perceived as more reliable, which reflects well on the company implementing the chatbot. Another participant suggested that the platform should include a way of providing feedback to the user in an effort to "close the loop" of communication. Ideally, this chatbot should be able to learn and adjust its responses in the case that it provides the user with poor feedback. The platform should also feature the chatbot's ability to learn from these experiences, so as not to make the same mistakes again in the future. It was likewise suggested that the platform should clearly define the boundaries of its functionality; for example, it should specify to the user at the beginning of a conversation that this particular chatbot is only capable of answering yes or no questions phrased by the user.

Still, other suggestions for the design of this platform incorporated a variety of participants' responses to keep the user engaged and interested in the conversation, including tips on its use of vocabulary, phrasing, dialog, etc. One user expressed a desire for the company to include a reference number for the conversation or alternate phone number at the beginning of a chat session, which would be extremely helpful in case of technical difficulties, such as a network disconnection or should the platform fail otherwise. Additionally, the company should ensure that users are not being required to provide their personal information a redundant number of times. This inconvenience is frustrating for users and may lead to a poor interaction that could potentially influence the user's decision to purchase services and/or products from a competing company in the future.

The main debate among the participants in this study was in whether or not it was more user-friendly to adopt a platform like Watson, which does not employ animation and voice functionality, or whether it was better to employ a platform such as Pandorabot or Verbot that would keep the user entertained and engaged with the technology. Demographically, professors were among those that consistently preferred a more professional-feeling experience with Watson and that found the use of avatars to promote a sense of amateurism in a company's technological capabilities. It was also said that the integration of an animated representation of the chat service detracted from the human-like aspect of the technology. There was no significant correlation to be found between undergraduate students versus graduate students in a user's preference of whether or not to include this feature.

3.3 System Usability Scale (SUS)

We calculated the SUS score of each of the participant's preferred chatbot platform so that we could quantify this data and make side-by-side comparisons. We adopted Brook's [3] equations to derive the numerical value of each user's individual chatbot session score. The equations we used to calculate this value are shown below:

For items 1, 3, 5, 7, 9:

$$\text{Sum1} = \text{score value} - 1 \tag{1}$$

For items 2, 4, 6, 8:

$$\text{Sum2} = 5 - \text{score value} \tag{2}$$

$$\text{SUS score} = 2.5 * (\text{sum1} + \text{sum2}) \tag{3}$$

Based on the values derived from this equation, we were able to compare each of these three platforms in terms of usability. This calculation provided us with a system to quantify usability as if it were an entity capable of being measured. The score of 68% has been used extensively in the literature as the usability threshold for interactive technologies (i.e., computer applications and websites). We used a more conservative threshold of 70%.

Combining each participant's individual SUS score and distributing these values in conjunction with their respective platforms, we were able to objectively quantify the usability of these three chatbot platforms. Individual participants scored the usability of Watson's platform with values of 100, 57.5, 77.5, 92.5, 75, 90, 80, and 82.5. Pandorabot was awarded individual usability scores of 92.5 and 85, which was representative of the two out of the ten total participants who chose it as their preferred platform overall. In order to compare Watson and Pandorabot's platforms to determine which of these chatbots the participants found to be more user-friendly, we had to combine the individual scores from each session to obtain overall numerical values. Overall, Watson scored 81.9 on the usability scale, while Pandorabot scored a value of 88.8. While Pandorabot achieved a better overall score, we must acknowledge that this value was calculated by taking only two participants' responses into consideration. Watson, on the other hand, earned its score based on the responses of a total of eight participants, which constituted the vast majority of our population of users. We therefore attribute this discrepancy to the statistical variation in these responses, given their population size.

5.0 Conclusion

A large telecommunication company provided us with the opportunity to recommend a chatbot technology and platform design that we believed the company should adopt to support internal communication within the organization. This technology would primarily serve to streamline communication between company's technicians in the field and engineers back at the facility. We approached this problem with the decision to design and conduct an original research study for the purpose of analyzing system usability with an emphasis on human factors engineering. A usability test was conducted to compare the usability of three chatbot platforms. For the purposes of this research study, we gathered feedback from ten participants, then proceeded to rate this feedback using a System Usability Scale (SUS). The results showed that overall, IBM's Watson was perceived to be the most user-friendly platform. Watson scored an average SUS score of 81.875 out of 100, while Pandorabot scored an 88.75 out of 100. Verbot was not assigned a SUS score due to the fact that not one of our ten participants opted in favor of this platform. While Pandorabot scored higher on the system usability scale, 80% of our participants preferred Watson's platform. The statistical variation among these participants' responses was attributed to a significantly small population of these participants choosing Pandorabot. The results suggest that IBM's Watson represents the technology best aligned with our human factors analysis. Watson had a perceived intelligence, a simplistic atmosphere, and was chosen by 80% of our participants.

The real-life application of a chatbot will save a company time and ultimately lead to financial gain because of the tasks it is able to take on and the ability to allow engineers to dedicate their time towards other tasks. As the intelligence and technology of chatbots evolve, chatbots will be able to take on more and more responsibilities.

References

[1] Copeland, By Michael. "The Difference Between AI, Machine Learning, and Deep Learning?" The Official NVIDIA Blog. N.p., 29 July 2016. Web. 06 Oct. 2016.

[2] "How Many Test Users in a Usability Study?" [Online]. Available: https://www.nngroup.com/articles/how-many-test-users/. [Accessed: 08-Jan-2017].

[3] Jordan, B. Thomas, I. L. McClelland, and B. Weerdmeester, *Usability Evaluation In Industry*. CRC Press, 1996.