

Towards an Empathetic Chatbot: Effects of a Chatbot Responding to User Frustration on Customer Service Interactions

Gerden Ibrahim

S4281608

Master Psychology: Behaviour Change

Research Paper

Internal Supervisor: Rob Bulterman

Company: Greenhouse Eindhoven

External Supervisor: Bas Ploeg

Radboud University Nijmegen

5th of July, 2019

Word count: 6973

Abstract

The purpose of the current study was to test whether an empathetic chatbot, which can recognize user frustration, would increase the number of words participants use during the interaction. Different studies have provided solutions on analyzing user emotion and other studies have provided insights into communication styles that are beneficial for chatbot interactions, however there is a lack of studies on a combination of matching communication to user emotion in the context of customer service bots. A software which analyzes frustration from written text was validated with a pilot study. This software was implemented into a new chatbot which sends either an agreeable response or mirrors users' punctuation style based on the level of frustration. 48 participants (19 male, 11 female, 1 not specified) with an average age of 23.4 played a frustration inducing game, after which they interacted either with the empathetic bot or a control bot. Results indicate that participants did not use more words during an interaction with the empathetic chatbot compared to the control bot, probably due to a lack of statistical power. Explorative analysis indicates that there is no difference on the intention to use the bot again, but that frustration was decreased more in users interacting with the empathetic bot.

Keywords: chatbot, frustration, emotion detection, customer service

To provide customers with efficient and innovative customer service channels, many companies nowadays have applied conversational agents to deal with customer's needs and wishes (Nordheim, 2018). Conversational agents (also known as *chatbots* or simply *bots*) are natural language interfaces which can respond to customers in real time (Brandtzaeg & Følstad, 2017). Compared to human agents they bring advantages of not getting tired, being available 24/7, and talking to a chatbot feels more natural than using a phone app (Brynjolfsson & McAfee, as cited by Nordheim, 2018). Dealing with (routine) customer requests means that there is less time and energy available to put into other, more creative tasks at the workplace. It can also be emotionally exhausting for employees when dealing with angry or stressful customers (Wang et al., 2013).

Chatbots increase chances of customers expressing honest feelings instead of responding with social desirability (Brandtzaeg & Følstad, 2017). When customers feel negatively during customer service interactions this feeling can be transcribed to the whole company. The study of Picard (2000) shows that customers who receive good support are more likely to use the services of the company again, even more so than customers who had no problems in the first place. Good support implies that the service team has been trained in

active listening and an appropriate use of empathy. Empathy describes the emotional response to someone's position. It communicates to the customer that the problem is understandable and acknowledged (Ickes, 1997). Picard (2000) underlines that these skills can also be of importance for computer interaction. Therefore, it is important to keep in mind that customer service agents need to be able to respond to human emotion.

Emotions (or emotional affect) are brief and specific human responses, for example anger or guilt (Gilovich, Keltner & Nisbett 2012). One type of emotional affect that is often experienced in the context of computer interaction, is frustration (Picard, 2000; Susskind, 2004). Frustration is defined as an emotional state that results from a need not being fulfilled (Lawson, as cited by Klein et al., 2002), which can, under circumstances, even lead to feelings of aggression (Gilovich et al., 2012).

Several studies have dealt with the concept of chatbots and emotion detection. Regarding emotion detection there is much discussion within the community on distinguishing between sentiment and emotional affect. Sentiment is defined as a classification of a positive versus negative valence and describes a general mood of a sentence. Emotional affect, on the other hand, describes more fine-grained emotions, like anger or happiness (Krcadinac, Pasquier, Jovanovic and Devedzic, 2013). Most studies focus on sentiment (e.g. Clavel & Callejas, 2016; Paltoglou, Gobron, Skowron, Thelwall & Thalmann, 2011; Priyadarshana et al., 2015), but Krcadinac et al., (2013) make the important point that, while affect and sentiment analysis are closely linked, analyzing emotional affect can give a more nuanced picture of the consumers' current state. Sentiment works better to describe an overall mood, but emotional affect describes feelings in one specific moment. For example, in a conversation a customer could feel frustrated after the first sentence, but angry after the second. The overall sentiment would be negative, but emotional affect is more specific to the individual sentences.

Studies on detection of emotional affect are less common and primarily based on physical responses. A classic variant is retrieving frustration levels by measuring participants' heart rate (Scheirer, Fernandez, Klein & Picard, 2002). Scheirer, Fernandez and Picard (1999) used expression glasses which can measure frustration based on eyebrow movement. Another option is a pressure mouse which can retrieve frustration from users' clicking behavior (Qi, Reynolds & Picard, 2001). These mechanisms might be promising, but they are not practical in daily communication as customers would first need to be provided with the tools, and this would add financial costs.

Retrieving frustration from written text would be a more practical solution as customers would not need any physical tools as the detection can be implemented online. However, research concerning frustration detection from written text, is limited. IBM released a program which can detect emotions from written text (Mostafa, Crick, Calderon & Oatley, 2016). The problem with programs like these is that they are based on online dictionaries that take context and sarcasm less into account. For example, a sentence like “My flight is delayed... amazing.”, is meant in a sarcastic manner, however, as the word amazing has a positive valence, the sentence would be scored as positive (Felbo, Mislove, Søggaard, Rahwan & Lehmann, 2017).

One of the most promising studies when it comes to emotion detection from written text is the study of Felbo et al., (2017), in which they used machine learning to analyze over 1246 million tweets. By combining tweets with their corresponding emoticons, they were able to create a tool (*deepmoji*) which translates written text into emoticons. Emoticons are symbolic representations often used in online communication that can be used to classify emotional contents of text (Felbo et al., 2017). By comparing a great number of tweets with the emoticon’s users use to support their tweets, a big dataset could be created that combines written text with emoticons. Instead of relying on definitions from dictionaries, the data is based on definitions individuals use in real life. The high number of tweets makes the overall estimation of emotional affect more reliable. In the earlier example (“My flight is delayed... amazing.”), the sentence would be combined with angry and annoyed emoticons, making the sarcastic nature of the input more visible.

While these studies provide interesting insights into the analysis of emotional affect of a conversation, they do not provide a solution on how to respond to the detected frustration. Different studies underline that demonstrating empathy in customer service interactions is important as it increases the chance that customers will use the services again (Picard et al., 2000), and that customers find it important to feel understood and acknowledged (e.g. Xu, Liu, Guo, Sinha & Akkiraju, 2017). The process behind this phenomenon can be explained with the Unified Theory of Acceptance and Use of Technology (UTAUT; Venkatesh, Morris, Davis, & Davis, 2003). According to UTAUT whether someone will use a new technology, as a chatbot, is predicted by the intention someone has about using the bot, or in other words whether they plan on choosing the bot as an option. The intention can be increased when someone feels positively about chatbots, also called positive attitude.

This positive attitude is predicted by the trust someone puts in the message of the customer service agent (Gremler Gwinner, & Brown, 2001) and the level of skepticism

someone has towards new technologies as chatbots (Araujo, 2018; Cornescu & Adam, 2013). Trust in this context means that customers find the message trustworthy and reliable. Trust predicts a positive attitude towards chatbots, which can be increased by creating a personal connection between human and bot (Gremler et al., 2001). A personal connection is defined as feeling a strong bond or affiliation towards someone else, often caused by seeing similarities between oneself and the communication partner (Duck, as cited by Gremler et al., 2001). Humans have the tendency to ascribe the same attributes to chatbots as to humans. This means that they interact with a bot in a similar way as they would with a human agent (Nass, Isbister & Lee, 2000). Following this logic, a personal connection could be created if consumers identify with the chatbot. Skepticism, on the other hand, should be decreased. Skepticism, in this context, describes a feeling of resistance towards innovation, a change of the satisfactory status quo (Araujo, 2018; Cornescu & Adam, 2013).

However, even if these factors are accounted for, intention does not necessarily lead to behavior. The relationship between intention and behavior often depends on whether the behavior stems from a conscious process (Sheeran, 2005). Which channel consumers choose when they have a customer service request is not always a conscious decision (Pieterse & Van Dijk, 2007). Therefore, it is important to keep automatic processes within consumers in mind. One automatic process that can influence channel choice is emotional affect. When a consumer is, for example, frustrated this can influence whether they will choose a chatbot or not. When negative affect, in this case frustration, is accounted for, the intention could lead to actual behavior meaning consumers will use the chatbot as an option for simple requests (Pieterse & Van Dijk, 2007).

Other studies have investigated which communication style used by chatbots consumers respond to. De Vries, Bakker-Pieper, Konings & Schouten (2011) have shown that communication style is often linked to personality. One personality trait that seems to be beneficial in customer service agents is *agreeableness*. Humans that score high on agreeableness tend to be less dominant, give in more easily and seem more empathetic (Herzig, Shmueli-Scheuer, Sandbank & Konopnicki, 2017). Communicating empathy is especially important within customer service where 40 percent of requests have an emotional nature (Xu et al., 2017). Herzig et al. (2017) showed that this can work in a chatbot scenario as well by implementing the communication style of agreeableness into a chatbot with the help of Natural Language Generation (NLG). NLG is a part of artificial intelligence, where a targeted communication style can be applied by training the model a big number of responses (Herzig et al., 2017). In more simple terms, NLG means that a bot can not only analyze text

but automatically produce it. In this study it was not possible to make use of NLG due to a time restriction and lack of expertise in artificial intelligence. Instead text is produced manually by a human writing a script.

When consumers do not feel negative affect, the liking principle of Cialdini (2012) can be a possibility to artificially create a personal connection between user and bot. Cialdini proposes that humans like others more if they resemble them. This liking principle can be increased by unconsciously mirroring the non-verbal behavior of the communication partner. This can increase perceptions of empathy, make someone seem more confident and persuasive (Pfeifer, Iacobini, Mazziotta & Dapretto, 2008; Van Swol, 2003). However, during an interaction between a human and a chatbot there is no possibility to observe nonverbal behavior (Kramer, 2007). Adjustments in communication style can bridge the gap between this: The consumers' use of punctuation can be mirrored. Different studies underline how punctuation can communicate different states and emotions. For example, ending a sentence with an exclamation mark instead of a period seems more friendly. (Hancock, Landrigan and Silver, 2007; Gunraj, Drumm-Hewitt, Dashow, Upadhyay & Klin, 2016). High frustration can therefore be responded to with an agreeable communication style, while low frustration can be responded to by mirroring interpunction.

To summarize, it can be said that there are multiple studies on frustration detection (Krcadinac et al., 2013; Scheirer et al., 1999; Scheirer et al., 2002; Qi et al., 2001) and emotional affect detection from written text (Felbo et al., 2017; Mostafa et al., 2016). Other studies have proven that being able to understand customer's emotions is important, especially in the context of customer service requests (Xu et al., 2017). Communicating agreeableness (Herzig et al., 2017; De Vries et al., 2011; Xu et al., 2017) and mirroring behavior can create a connection between human and bot (Cialdini, 2012; Hancock et al., 2007; Gunraj et al., 2016; Pfeifer et al., 2008; Van Swol, 2003). However, an important limitation is that there is no research done on the combination: A chatbot which can detect frustration and respond in a way that will make consumers choose the chatbot as an option.

Therefore, the question of this study will be: What is the influence of an empathetic chatbot on choosing the chatbot for simple customer service requests? Empathetic in this context means that the chatbot can detect the consumer's level of frustration and match its communication style accordingly. The target is limited to simple requests as chances are too low that someone will use a channel that they are not familiar with for more complex questions (Pieterse & Van Dijk, 2007).

Based on the research question the following hypothesis was set up: Consumers will repeatedly use the chatbot which detects frustration and changes its communication style based on the customer's level of frustration. This will be measured by the average number of words participants use during an interaction with the chatbot. Due to a time restriction it is not possible to observe follow-up behavior. Users with a positive affect subconsciously use more words in online communication (Hancock et al., 2007). When positive affect is accounted for, chances are higher that someone will use the chatbot again (Sheeran, 2005). Therefore, it was decided to use number of words as a predictor for future behavior. Next to this experimental measure, two additional variables will be explored. It is expected that consumers which interact with the empathetic chatbot will show a decrease of frustration. The third hypothesis is that participants that interact with the empathetic chatbot will have a higher intention to use the chatbot again.

In collaboration with Greenhouse Eindhoven a tool was created which can detect the degree of frustration from written text. This tool is implemented into a chatbot which was created specifically for this study: Claire the empathetic chatbot. To create a frustrating scenario, participants will play a frustrating smartphone game. They will be asked to give a review to Claire the chatbot. Based on the level of frustration in one message, participants either receive an agreeable response (to high frustration) or their punctuation is mirrored (for low frustration).

The target group will consist of consumers aged 18 to 35 as chances are higher that individual in this specific age group will be early adopters when it comes to technology. This does not necessarily mean that anyone in this age group is an early adopter, however individuals within this range often grew up with computers and could therefore be more risk-oriented and curious about new technologies (Williams & Page, 2011). When developing new technologies chances are higher that they will accept and understand them first (Brandtzaeg & Følstad, 2017). When proven successful with this group, it can be expanded to other age groups.

If the chatbot proves to be successful it can be used to answer simple customer service requests within companies and organizations. This way employees would have more time to deal with more important tasks which can increase creativity and motivation (Wang et al., 2013). It could also be used to detect frustration in the first line of customer service where questions are being filtered: When customers show a high level of frustration, they could be connected with an employee that is better at dealing with frustration. Lastly, the chatbot could deal with immediate questions when there are no employees available. This way companies

can make sure that there is always an option where their consumers can let off steam or be helped immediately in an empathetic way.

Method

To see whether participants prefer the chatbot which can detect frustration and change its responses depending on the level of frustration observed in a single message, a frustration increasing scenario was created, after which they answered questions about the game to a chatbot. Frustration within the conversation was analyzed by a frustration analysis tool which was validated with a pilot study.

Pilot (Study 1)

The purpose of the pilot study was to test whether frustration can be correctly estimated with the frustration analysis tool which was created within Greenhouse. The hypothesis is that there should be a positive correlation between the frustration scores created by the frustration analysis tool and the subjective frustration scores of participants.

Participants

In total 22 participants took part in the pilot study. One participant had to be removed from the dataset for not giving consent and one outlier were removed, leaving 21 participants. Participants were recruited within Greenhouse ($N = 10$) as well as from social circles ($N = 11$). To reach a power of .80 with a small effect size ($\alpha = .10$), a sample size of 58 participants would have been needed. Because of the time restriction it was decided to go on with the results nonetheless the low power. It is advised to repeat this test with a higher number of participants.

Materials and Procedure

Materials consisted of a smartphone game application, a survey, a chatbot and a frustration analysis tool.

Game. The smartphone version of an already existing game was used (www.theworlds-hardestgame.com). The game was originally designed to be very hard to beat, possibly increasing frustration. To mask the purpose of the study, participants were asked to test a game app and give a review about the game to a chatbot.

Survey. With the help of the survey program *Qualtrics* (www.qualtrics.com) a survey was created. The survey consisted of a description of the purpose of the study, an informed consent form and three closed questions. Based on the study of Klein et al., (2002),

participants were asked how frustrated they think they got during the game, scored on a 10-point scale. Two additional questions were asked to mask the real purpose of the study (see appendix A for pilot survey).

Chatbot and analysis tool. To retrieve textual information from participants a chatbot was created. A conversation script about the users gaming experience was manually typed and this was implemented into a chatbot (see appendix C for conversation script). The bot therefore asked the prepared questions to each participant. This was done in the same way for every participant, which means everyone received the same questions and responses in the same order. A frustration analysis tool was used to analyze the textual data that was received from the chatbot conversation. The tool was created within Greenhouse and was based on the emotion detection tool *deepmoji* but translated the emoticons into scores. The concrete concept behind the tool cannot be elaborated fully here as the focus of this project was on the psychological and communication processes. The chatbot and analysis tool were not combined for the pilot due to a lack of time. Instead textual information that was retrieved from the conversations was manually put into the analysis tool. Lastly participants were debriefed and asked what they thought the purpose of the study was. No participant mentioned knowing that the study was about frustration.

Measurement

The input that participants gave during the conversation with the chatbot were analyzed by the frustration analysis tool which was created by Greenhouse. One outlier was removed. For each participant average frustration percentages were created which were converted into decimal scores. The chatbot frustration scores and the subjective frustration of participants were transferred into SPSS 21. To test whether there is a positive association between subjective frustration scores and chatbot frustration scores, Pearson correlations were calculated.

Results pilot study

The one-tailed Pearson correlation indicates that the subjective frustration scores and the chatbot frustration scores are positively correlated, $r(19) = .43, p = .03$. The results suggest that there is a relatively high, positive association between the frustration scores produced by the analysis tool and the frustration scores participants gave themselves. This could mean that the frustration tool can recognize frustration on a level that is comparable to

self-reports on frustration. However, as mentioned before these results should be dealt with carefully due to the small sample size.

Effect study (Study 2)

Participants

In total 53 participants took part in the study. Five participants had to be removed from the dataset because of technical issues, leaving 48 participants with an average age of 23.4 ($SD = 3.1$). Out of the remaining 48 participants, 28 were male, 19 were female and 1 did not specify. One part of the participants was recruited within Greenhouse and personal circles and participated voluntarily ($N = 38$). The other part was recruited via the online participation system of Radboud University Nijmegen and was rewarded with course credit ($N = 10$).

Design

The current study used a between-subject design with type of chatbot (empathetic Claire vs. control Claire) as a between-subject factor and the subjective frustration score before the chat as a covariate. The dependent variable was the average number of words participants used during the interaction. From the script nine questions were chosen. Only responses to those nine questions were used for analysis. The reason for this is that the first questions of the script were neutral, control questions, asking participants for their participant numbers (see appendix C).

Materials and Procedure

The materials for the effect study consisted of a survey, the same smartphone game used during the pilot study and two chatbots, one used for the experimental group and one used for the control group.

Survey and game. The survey was created with Qualtrics and looked similar to the survey used during the pilot. The survey consisted of a description of the study, an informed consent form, the same three questions about the game used in the pilot (subjective frustration estimation after the game and the two masking questions). The game was not implemented into the survey but was played on a smartphone. Participants were asked to play for exactly five minutes after which they heard a timer and had to stop playing. After the game participants were asked whether they would use the chatbot again. This intention measurement was scored in percentages. Next, a follow-up frustration question was asked and one additional masking question. Besides the intention measurement, all questions were

scored on a 10-point scale. Lastly, participants filled in demographic information (see appendix B for effect study survey).

Chatbots. Participants were randomly assigned to one of two groups. Group one interacted with Claire, the empathetic chatbot; Group two interacted with Claire, the control chatbot. To make sure a possible effect would not arise from the name of the chatbot, both chatbots received the same name. The chatbot used for the pilot study was used as a base to create the control and the experimental chatbots. For the control chatbot, the adjustment was made that it did not only ask questions (as in the pilot) but also added a standard response to each answer. In the pilot study the purpose of the chatbot was mainly receiving textual data, in the effect study the purpose was testing the abilities of the bot during an interaction that is similar to a customer service interaction. Control Claire gives the same responses to each participant regardless of the emotional affect of the sentence.

The chatbot used for the experimental condition makes a frustration estimation for each response sent by the participant. Based on the level of frustration, one out of two responses are chosen. For each sentence a frustration score on a scale from 0 to 10 is created. The frustration cut-off scores in the tool is by default five as this is the cut-off for the probability of either high or low frustration. If the score of the response equals 5 or higher, the chatbot sends a response that was created in an agreeable communication style. To create agreeable responses the *Communication Styles Inventory* (CSI) used in the study of Herzig et al., 2017 was used. Standard responses used for the script of the control chatbot were adjusted to communicate an agreeable style. One example of an agreeable response would be: *I'm sorry... I wish you had a better experience! I would still like to ask you a couple more questions if that's okay?* (see appendix C for the complete script).

If a score lower than 5 is detected, a low frustration response is sent. For low frustration sentences, the same responses used in the control condition are used with one minor adjustment: For a low frustration response, the chatbot program first analyzes whether an exclamation mark is used; If an exclamation mark is used, it is copied in the next following response of the chatbot. When there is no exclamation mark used, the chatbot ends the response with a period. Any other type of punctuation besides an exclamation mark or a period could change the syntax of the sentence, thus only exclamation marks are mirrored. The responses for both chatbots were automated, which means that the rules for responses were implemented into the code of the chatbot and did not have to be added manually. Lastly participants were debriefed and asked what they thought the purpose of the study was. No participant mentioned realizing that the study dealt with frustration.

Data-analysis

For each participant the number of words used in answering the earlier mentioned set of nine questions were used. The numbers used in the different sentences were added up for each participant to receive a total score of words and then inserted into the dataset. A one-way ANCOVA was conducted with the average number of words as the dependent variable, type of chatbot as a between-subject factor and subjective frustration score before the chatbot conversation as a covariate. The covariate was added to control for possible influence by frustration caused by the game.

To explore whether there is a difference in frustration reduction between the two conditions a mixed ANOVA was conducted with type of chatbot (empathetic Claire versus control Claire) as a between-subject factor, time of frustration measurement (before and after the chat) as a within-subject factor and level of frustration as the dependent variable. To explore whether participants in the experimental condition rated the chance higher of using the chatbot again, an independent-samples t-test was conducted with intention as the dependent variable and type of chatbot as a between-subject factor.

Results effect study

Number of words. To test whether the average number of words used in the conversation with the chatbots were matched between the two groups, a one-way ANCOVA was conducted with type of bot (empathetic vs. control) as a between-subject factor, the level of frustration after the game as a covariate and the average number of words used during the conversation as the dependent variable. The results of the one-way ANCOVA indicate that there is no significant effect of level of frustration on the average number of words, $F(1, 43) = .54, p = .47, \eta^2 = .01$. This means that level of frustration does not predict the number of words someone uses during the interaction. The main effect of type of bot was not significant when controlled for level of frustration, $F(1, 43) = .01, p = .91, \eta^2 = .00$. The means between the two groups are equal ($M_{\text{empathetic}} = 117.88, SD_{\text{empathetic}} = 63.58, M_{\text{control}} = 114.48, SD_{\text{control}} = 61.75$). This indicates that there is no difference in the average number of words used during the interaction between the two conditions when controlled for the level of frustration of participants. The lack of power ($1-\beta = .05$), leads to a number of implications for the conclusion which can be found in the discussion section.

Frustration decrease. To explore whether there is a difference in frustration decrease between the two conditions before and after the chat with the bot, a mixed ANOVA was

conducted. Results indicate that the main effect of time of measurement is significant, $F(1, 45) = 68.83, p > .001, \eta^2 = .61$. Figure 1 illustrates that there seems to be an overall decrease in frustration for both groups over time. The main effect for type of chatbot is also significant, $F(1, 45) = 4.75, p = .04, \eta^2 = .1$. This indicates that there seems to be a difference in frustration decrease between the two conditions. Looking at the interaction effect between time of measurement and type of chatbot, a significant interaction effect is observed, $F(1, 45) = 5.47, p = .02, \eta^2 = .11$. The interaction effect indicates that, although frustration decreased in both groups, there is a significant difference in frustration decrease for both groups when controlled for time of measurement. This suggests that both groups seemed to feel less frustrated after the interaction with the chatbot, but that this effect seems to be stronger for participants who interacted with the empathetic chatbot (see figure 1). It should be noted that the post-hoc observed power was not very high ($1-\beta = .63$), possibly due to the small sample size. This needs to be kept in mind when interpreting the results. See discussion for implications.

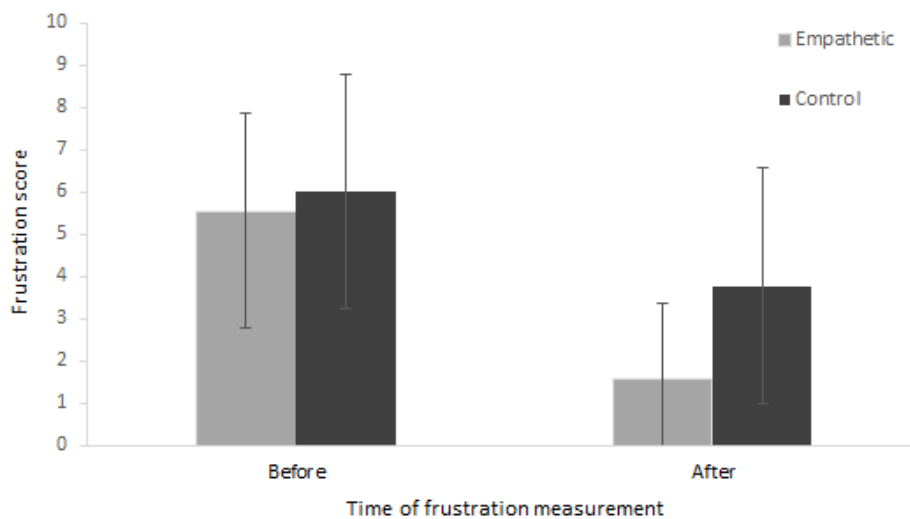


Figure 1: Frustration scores before and after the chatbot conversation between the empathetic chatbot and the control chatbot. Standard deviations are represented in the figure by the error bars attached to each column.

Intention to use the chatbot again. The explorative independent-samples t-test with type of chatbot (empathetic vs. control) on the intention to use the chatbot again shows that there is no significant difference between the two conditions, $F(1, 45) = .12, p = .89$. This indicates that the intention to use the chatbot again is approximately the same for both the

empathetic condition ($M_{\text{empathetic}} = 63.42$, $SD_{\text{empathetic}} = 20.52$) as well as for the control condition ($M_{\text{control}} = 62.48$, $SD_{\text{control}} = 25.12$).

Discussion

The aim of the current study was to answer the question whether interacting with an empathetic chatbot can influence the number of words used during that interaction. Empathetic in this context means that the level of frustration within a sentence can be automatically estimated after which a response is given that is personalized to the degree of frustration of that sentence. Furthermore, it was explored whether the empathetic chatbot can decrease feelings of frustration (hypothesis 2) and whether participants interacting with the empathetic chatbot would have a higher intention of using the bot again (hypothesis 3).

To test whether frustration within sentences can be correctly estimated by the emotion analysis tool, a pilot study was set up. With this pilot study it was validated that the frustration score that the bot creates is in line with participants' self-reported frustration scores. It should be noted, that these results should be handled with caution due to the rather small sample size, due to a shortage of time. Future research could replicate the study to get a more accurate picture of the analysis tool.

Regarding the first research hypothesis whether participants interacting with the empathetic chatbot use more words in their responses than participants interacting with the control chatbot, the results indicate that there is no significant difference between the average number of words used in the two conditions, when controlled for the level of frustration. From this it can be concluded that interacting with the empathetic chatbot does not increase the number of words used during the conversation and that this does not depend on the user's level of frustration. This results in different theoretical and methodological implications for the first research hypothesis.

Looking at the conversations it appears that, in most cases, the chatbot can distinguish between messages high or respectively low in frustration. When it comes to more vague sentences, the estimation is not always correct (e.g. *I felt as if I could have played longer*). This is, however, based on subjective observations as there was no possibility to see which emotion participants were really feeling during the conversation. A possible explanation could be that participants are feeling more sets of emotions, especially due to the experimental setting (Rosenthal-von der Pütten, Krämer, Hoffmann, Sobieraj & Eimler, 2012). The theory behind the responses sent by the empathetic chatbot is focused solely on

frustration, which makes it difficult to properly respond to participants when they are feeling other emotions. It was aimed to control for this factor by creating a frustrating situation but there can always be individual differences in how individuals interpret a game (Siemer, Mauss & Gross, 2007).

Another methodological explanation for the results is the lack of power, possible due to the small sample size. In the current study, despite large efforts, it was not possible to recruit more participants as the project was limited to a specific time span. The small sample size results in several consequences for the found results.

Concerning the first hypothesis, the small sample size could be a reason for the dependent variable to not be very reliable. The dependent variable was inspired by the study of Hancock et al., (2007), where it was concluded that positive affect and increase in word use cohere. As subjective measurements are not always reliable, this implicit behavior measurement was chosen that could give an indication on the long-term use of chatbots (Hancock et al., 2007, Venkatesh et al., 2013). Hill, Randolph and Farreras (2015) found that users send 4.29 words on average per message when interacting with a chatbot (this is opposed to 7.95 words in a human to human conversation). In the current study, the average number of words used in one message is 13.1 words for the empathetic bot condition and 12.72 words for the control condition. As the average number of words in each message was not part of the analysis, no concrete conclusions can be drawn. However, it seems as if the number of words used by participants in the current study is higher on average compared to their results. This could indicate that interacting with a chatbot in an experimental setting increased the positive affect for both conditions, demonstrated by the number of words. It would be interesting for future research to put more emphasis on single messages than on whole conversations. Instead of analyzing a whole conversation, users' responses after each sentence could be measured to see whether an empathetic response directly increases the number of words in the users' response.

Looking at the theory we see some differences between the current study and other research dealing with comparable thematics. In the current study, exclamation marks used in messages with a low degree of frustration were mirrored by the chatbot. However, when looking at the different input from conversations, it was observed that exclamation marks were scarcely used. Looking at combinations of different studies, mirroring users' punctuation could increase liking, which could increase positive affect (Cialdini, 2012; Pfeifer et al., 2008, Van Swol, 2003). In general, mirroring exclamation marks might have been too subtle to have an effect on participants. An improvement could be to mirror

punctuation in general, and not limit the mirroring to exclamation marks. It could also be tested if other - subtle - individual differences in communication style could be copied. For example, in the current study some participants started sentences with uppercase letters and others with lowercase letters. It could be interesting to see whether mirroring the participants' use of letter case could increase positive affect.

Since responses were adjusted for each sentence and not for each participant, conclusion about results can only be drawn on a combination of high and low frustration responses. To make this clearer: During the interaction with the chatbot, each participant received eleven questions about the gaming experience, and responded to them. For example, participants were asked for their general opinion but also more specific for their opinion on the music. A participant could have enjoyed the game, but still be frustrated by the music. The frustration score would be low in the beginning of the interaction but could increase when answering the question about the music. The current results concerning the average number of words could mean two things: On the one hand it is possible that users interacting with the empathetic chatbot did not feel more positively about the interactions. Another possible explanation is that number of words might not have been representative for feelings of a positive affect in this study.

In the current study, number of words used during the interaction with the chatbot was chosen as an indirect measurement of positive affect. When looking at the two conditions it seems as if there is no difference in positive affect between users interacting with the empathetic or the chatbot, respectively. This is not in line with results from previous studies. Herzig et al. (2017) demonstrated that users prefer an agreeable communication style when interacting with a chatbot. This agreeable communication style should seem more empathetic, leading to a more positive affect from the user towards the chatbot (Xu et al., 2017). Xu et al., (2017) and Klein et al., (2002) showed that an empathetic chatbot is especially wished for when customers feel highly emotional. In the current study the empathetic chatbot did not increase the number of words used, which could indicate that positive affect was not higher in the condition where participants interacted with the empathetic chatbot opposed to the control bot. However, results of the second hypothesis suggest that there is a stronger decrease in frustration in the empathetic condition, which is discussed in more detail in the following paragraph. This does not necessarily suggest that positive affect is high, but it does seem as if participants feel better after interacting with the empathetic bot. For future research, replacing the dependent variable of number of words with another behavior measurement, as follow-up usage of the chatbot, could give more insight into the discrepancies found.

In anticipation that the dependent variable of number of words might not be representative enough, it was chosen to explore whether the empathetic chatbot could decrease the level of frustration users are feeling (hypothesis 2). Looking at the results of this explorative analysis, we can see that there is a difference in frustration decrease between the two groups: Participants interacting with the empathetic chatbot have a significantly higher decrease in frustration levels than participants interacting with the control chatbot. This indicates that interacting with the empathetic bot can decrease frustration, even when controlled for time. Again, it should be noted that the power of these results is only moderately high. More research with higher sample sizes is needed to get more correct estimations for the population. However, when comparing this to the results of Herzig et al., (2017), Xu et al., (2017) and Hancock et al., (2007) it seems like the empathetic chatbot can make participants feel more positively or relieve frustration to a certain degree. It therefore seems more plausible that number of words was not representative enough in the current study.

Lastly, it was explored whether participants interacting with the empathetic chatbot would have a higher intention to use the chatbot again compared to participants interacting with the control chatbot (hypothesis 3). The results show that there is no difference between the intention to use the chatbot again for both conditions. This means that someone interacting with the empathetic chatbot does not estimate the chance of using the chatbot higher than someone who interacted with the control bot. The results indicate that intention is moderately high for both groups. Several possibilities could give more insights into why intention was moderately high in both conditions, despite the difference in chatbots. One explanation could be the lack of alternatives. Participants were asked whether they would use the chatbot again if they had to answer follow-up questions about their gaming experience. However, it was not stated whether an alternative would be a face to face conversation, a phone call, a questionnaire or a different chatbot. The uncertainty might have been a reason for participants to estimate the chance rather high to use the chatbot again, despite what participants were thinking about that specific chatbot. Literature on uncertainty in the context of chatbots is limited but Sanglé-Ferrière and Voyer (2017) investigated the role of uncertainty and customer service chat interactions with a human agent, where it was concluded that customers prefer chat when felt risk is high, for example when they want to be in control over the course of the interaction or do not want to be influenced by a human due to the distance that chat provides. Even though the mentioned study did not specifically deal with chatbots, similar feelings might arise when thinking about the option of a chatbot.

Furthermore, most participants were young and male. Research shows that this particular group might have a preference for chatbots in general (Wang & Wang, 2010).

Next to the methodological and theoretical alternative explanations for the found results, the study had some limitations. First, the current study was conducted in an experimental laboratory setting. In real life situations, individuals might respond differently to a chatbot. The aim was to make the situation as personally relevant as possible to participants by letting them play a game and ask for their personal opinions, however individuals might make different choices in real life settings. Second, as mentioned before, the current study only focused on frustration as an emotion. Further research is needed to estimate more complete sets of emotion.

Despite these few limitations the results of this study could be interesting for several different sectors. First, the frustration analysis tool can be used to receive an objective alternative to analyze text data as the tool is based on a big dataset. The tool could also be used in the first line of customer service. When high frustration is detected, a human agent that is trained in dealing with strong emotions could be connected to the customer. It was also demonstrated that the empathetic chatbot might be able to decrease frustration. The empathetic chatbot could therefore be used as an option to add to customer service channels where customers would receive immediate support and still feel appreciated due to the empathetic nature of the chatbot. In other sectors, such as health, the chatbot could function as a way for patients to communicate even when there is no one around but still feel acknowledged. In general, more research is needed to create a chatbot that will make users feel understood on an interpersonal level. Future research should focus on the analysis of a varied set of emotions and the way these can be matched with corresponding personalized communication styles. This could eventually result in a chatbot that is able to copy the human concept of empathy. However, ethical remarks need to stay in focus. The goal of the empathetic chatbot is supporting the daily lives of humans and make them easier, not replace them.

References

- Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183-189.
doi:10.1016/j.chb.2018.03.051
- Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. *Internet Science*, 377-392.
doi:10.1007/978-3-319-70284-1_30
- Business Insider Intelligence. (2016, December 14). 80% of businesses want chatbots by 2020. Retrieved from <https://www.businessinsider.com/80-of-businesses-want-chatbots-by-2020-2016-12?international=true&r=US&IR=T>
- Cialdini, R. B. (2012). Liking: The friendly thief. In *Influence: The Psychology of Persuasion* (2nd ed., pp. 126-156). New York, NY: HarperCollins.
- Clavel, C., & Callejas, Z. (2016). Sentiment Analysis: From Opinion Mining to Human-Agent Interaction. *IEEE Transactions on Affective Computing*, *7*, 74-93.
doi:10.1109/taffc.2015.2444846
- Cornescu, V., & Adam, C. (2013). The consumer resistance behavior towards innovation. *Procedia Economics and Finance*, *6*, 457-465. doi:10.1016/s2212-5671(13)00163-9
- De Vries, R. E., Bakker-Pieper, A., Konings, F. E., & Schouten, B. (2011). The communication styles inventory. *Communication Research*, *40*, 506-532.
doi:10.1177/0093650211413571
- Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. doi:10.18653/v1/d17-1169
- Gilovich, T., Keltner, D., Chen, S., & Nisbett, R. E. (2012). Emotion. In *Social Psychology* (3rd ed., pp. 194-230). New York, NY: W. W. Norton.
- Gremler, D. D., Gwinner, K. P., & Brown, S. W. (2001). Generating positive word-of-mouth communication through customer-employee relationships. *International Journal of Service Industry Management*, *12*, 44-59. doi:10.1108/09564230110382763
- Gunraj, D. N., Drumm-Hewitt, A. M., Dashow, E. M., Upadhyay, S. S., & Klin, C. M. (2016). Texting insincerely: The role of the period in text messaging. *Computers in Human Behavior*, *55*, 1067-1075. doi:10.1016/j.chb.2015.11.003

- Hancock, J. T., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication. *Proceedings of the SIGCHI conference on Human factors in computing systems*. doi:10.1145/1240624.1240764
- Herzig, J., Shmueli-Scheuer, M., Sandbank, T., & Konopnicki, D. (2017). Neural response generation for customer service based on personality traits. *Proceedings of the 10th International Conference on Natural Language Generation*. doi:10.18653/v1/w17-3541
- Hill, J., Randolph Ford, W., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human-human online conversations and human-chatbot conversations. *Computers in Human Behavior*, 49, 245-250. doi:10.1016/j.chb.2015.02.026
- Ickes, W. J. (1997). *Empathic accuracy*. New York, NY: Guilford Press.
- Klein, J., Moon, Y., & Picard, R. (2002). This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14, 119-140. doi:10.1016/s0953-5438(01)00053-4
- Krcadinac, U., Pasquier, P., Jovanovic, J., & Devedzic, V. (2013). Synesketch: An open source library for sentence-based emotion recognition. *IEEE Transactions on Affective Computing*, 4, 312-325. doi:10.1109/t-affc.2013.18
- Mostafa, M., Crick, T., Calderon, A. C., & Oatley, G. (2016). Incorporating Emotion and Personality-Based Analysis in User-Centered Modelling. *Research and Development in Intelligent Systems XXXIII*, 383-389. doi:10.1007/978-3-319-47175-4_29
- Nass, C., Isbister, K., & Lee, E. J. (2000). Truth is beauty: Researching embodied conversational agents. *Embodied conversational agents*, 40, 374-402. doi:10.7551/mitpress/2697.003.0016
- Nordheim, C. B., (2018). *Trust in chatbots for customer service: Findings from a questionnaire study (thesis)*. Retrieved from <https://www.duo.uio.no/>
- Paltoglou, G., Gobron, S., Skowron, M., Thelwall, M., & Thalmann, D. (2010). Sentiment analysis of informal textual communication in cyberspace. *In Proc. Engage 2010, Springer LNCS State-of-the-Art Survey*, 13-25. Retrieved from <http://www.ofai.at/>
- Pfeifer, J. H., Iacoboni, M., Mazziotta, J. C., & Dapretto, M. (2008). Mirroring others' emotions relates to empathy and interpersonal competence in children. *NeuroImage*, 39, 2076-2085. doi:10.1016/j.neuroimage.2007.10.032
- Picard, R. W. (2000). Toward computers that recognize and respond to user emotion. *IBM Systems Journal*, 39, 705-719. doi:10.1147/sj.393.0705

- Pieterse, W., & Van Dijk, J. (2007). Channel choice determinants; an exploration of the factors that determine the choice of a service channel in citizen initiated contacts. *Proceedings of the 8th annual international conference on digital government research: bridging disciplines & domains*, 173-182. doi:10.3990/1.9789036528078
- Priyadarshana, Y., Gunathunga, K., Perera, K. N., Ranathunga, L., Karunaratne, P., & Thanthriwatta, T. (2015). Sentiment analysis: Measuring sentiment strength of call centre conversations. *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. doi:10.1109/icecct.2015.7226053
- Qi, Y., Reynolds, C., & Picard, R. W. (2001). The Bayes Point Machine for computer-user frustration detection via pressuremouse. *Proceedings of the 2001 workshop on Perceptive user interfaces - PUI '01*. doi:10.1145/971478.971495
- Qualtrics. (2019). *Qualtrics*. Retrieved from <https://www.qualtrics.com>
- Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2012). An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics*, 5, 17-34. doi:10.1007/s12369-012-0173-8
- Sanglé-Ferrière, M., & Voyer, B. (2017). Understanding chat perceptions in a customer assistance channel. *ACR North American Advances*, 45, 862-864. Retrieved from <http://www.acrwebsite.org/volumes/1024309/volumes/v45/NA-45>
- Scheirer, J., Fernandez, R., & Picard, R. W. (1999). Expression glasses. *CHI '99 extended abstracts on Human factors in computing systems - CHI '99*. doi:10.1145/632716.632878
- Scheirer, J., Fernandez, R., Klein, J., & Picard, R. W. (2002). Frustrating the user on purpose: a step toward building an affective computer. *Interacting with Computers*, 14, 93-118. doi:10.1016/s0953-5438(01)00059-5
- Sheeran, P. (2005). Intention-behavior relations: A conceptual and empirical review. *European Review of Social Psychology*, 12, 1-36. doi:10.1002/0470013478.ch1
- Siemer, M., Mauss, I., & Gross, J. J. (2007). Same situation--Different emotions: How appraisals shape our emotions. *Emotion*, 7, 592-600. doi:10.1037/1528-3542.7.3.592
- Susskind, A. M. (2004). Consumer frustration in the customer-server exchange: The role of attitudes toward complaining and information inadequacy related to service failures. *Journal of Hospitality & Tourism Research*, 28, 21-43. doi:10.1177/1096348003257328

- Van Swol, L. M. (2003). The effects of nonverbal mirroring on perceived persuasiveness, agreement with an imitator, and reciprocity in a group discussion. *Communication Research, 30*, 461-480. doi:10.1177/0093650203253318
- Venkatesh, Thong, & Xu. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly, 36*, 157-178. doi:10.2307/41410412
- Wang, M., Liu, S., Liao, H., Gong, Y., Kammeyer-Mueller, J., & Shi, J. (2013). Can't get it out of my mind: Employee rumination after customer mistreatment and negative mood in the next morning. *Journal of Applied Psychology, 98*, 989-1004. doi:10.1037/a0033656
- Wang, H., & Wang, S. (2010). User acceptance of mobile internet based on the Unified Theory of Acceptance and Use of Technology: Investigating the determinants and gender differences. *Social Behavior and Personality: An International Journal, 38*, 415-426. doi:10.2224/sbp.2010.38.3.415
- Williams, K. C., & Page, R. A. (2011). Marketing to the generations. *Journal of Behavioral Studies in Business, 3*, 37-53. Retrieved from <http://www.aabri.com/jbsb.html>
- Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on social media. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3025453.3025496

Appendix A: Questions from pilot survey

Explanation of the game

I want to know how well playing a desktop game on a smartphone works. For that purpose, I want to ask you to play a game that I picked out. The game is pretty self-explanatory. When you finish a level, you can just click on the next one. You have 5 minutes to get as far as you can. After 5 minutes a timer will ring after which I want to ask you to immediately stop playing (even if you aren't finished yet). After the game I would like to hear your opinions on the game and on the device itself.

Questions about the game

Now that you have finished playing the game, I have a few questions about your experiences with it.

On a scale from 0 to 10 how well do you think this game works on a smartphone? (0 = Not good at all;

10 = This is the best way to play it)

On a scale from 0 to 10, how frustrated do you think you got during the game?

(0 = Absolutely not frustrated at all; 10 = The most frustrated I have ever been in my life playing a game!)

Overall how much fun did you have playing the game?

(0 = No fun at all; 10 = The most fun I have ever had playing a game in my life)

Appendix B: Questions from effect study survey

Explanation of the game

I want to know how well playing a desktop game on a smartphone works. For that purpose, I want to ask you to play a game that I picked out. The game is pretty self-explanatory. When you finish a level, you can just click on the next one. You have 5 minutes to get as far as you can. After 5 minutes a timer will ring after which I want to ask you to immediately stop playing (even if you aren't finished yet). After the game I would like to hear your opinions on the game and on the device itself.

Questions about the game

Now that you have finished playing the game, I have a few questions about your experiences with it.

On a scale from 0 to 10 how well do you think this game works on a smartphone? (0 = Not good at all;

10 = This is the best way to play it)

On a scale from 0 to 10, how frustrated do you think you got during the game?

(0 = Absolutely not frustrated at all; 10 = The most frustrated I have ever been in my life playing a game!)

Overall how much fun did you have playing the game?

(0 = No fun at all; 10 = The most fun I have ever had playing a game in my life)

Instructions for chatbot

Thank you for filling out the questions! Now we would like to hear your opinion in more detail, but we want to do that in a more anonymous way. We created a chatbot which is going to ask you a couple of questions on the game. Imagine this is a regular conversation. There are no right or wrong answers, we just want to hear your honest opinion. The more information you provide the better! That helps us getting a varied image.

In the beginning of the interaction the chatbot will ask for the participation number that you received in the beginning of the study. This way we can make sure that the conversation stays anonymous.

Follow-up questions:

Imagine you had to answer some follow up questions on your gaming experience. What are the chances that you would choose to chat with Claire for that purpose?

(0% = I would definitely not choose Claire; 100% = I would most definitely choose Claire)

This is almost the end of the study. Now that you had to give the review, I would like to know how you are feeling at this particular moment.

On a scale from 0 to 10, how frustrated do you feel at this moment?

(0 = Absolutely not frustrated at all; 10 = The most frustrated I have ever been in my life!)

How happy are you at this moment?

(0 = Not happy at all; 10 = The happiest I have ever been in my life)

Demographic information:

What is your gender?

- Female
- Male
- Other

What is your age? _____.

Appendix C: Script for chatbot interactions

Responses for the control chatbot, responses for low frustration messages, responses for high frustration messages. Only sentences in bold are used for analysis.

1. Hi there! Before we start talking about the game, I was wondering what the participant number is that you received?
2. Is participant number your participant number?

If yes, go to question 3.

If no, follow-up question: Enter your participant number please. (repeat question 2)

3. **Great! What did you think of the game in general?**

Control response: Okay thanks.

Low frustration response: Okay thanks (+ interpunction)

High frustration response: Oh that sounds somewhat frustrating. I hope it didn't bother you too much...

4. **Can you tell me how the game made you feel? Feel free to add as much detail as you like.**

Control response: Thanks for the general feedback, now let's talk about the different parts of the game.

Low frustration response: Thanks for the general feedback, now let's talk about the different parts of the game (+ interpunction)

High frustration response: I'm sorry.. I wish you had a better experience! I would still like to ask you a couple more questions if that's okay?

5. **What did you think about playing it on a phone?**

Control Response: Got it.

Low frustration response: Got it (+ interpunction)

High frustration response: Being a bot I can't really judge, but what you are saying sounds very plausible.

6. **What did you think about the music in the game?**

Control response: Okay thank you, I have another question.

Low frustration response: Okay thank you, I have another question (+ interpunction)

High frustration response: Yeah that makes sense... It might have been a bold choice.

7. **What did you think about having to stop after 5 minutes?**

Control response: Noted.

Low frustration response: Noted (+ interpunction)

High frustration response: Oh no.. I hope you still enjoyed it a little bit.

8. **How far did you get in the game?**

Control response: Okay.

Low frustration response: Okay (+ interpunction)

High frustration response: Oh well, it' still quite impressive I would say.

9. Did you use any strategies in the game? If yes, what were they?

Control response: I guess everyone has different strategies.

Low frustration response: I guess everyone has different strategies (+interpunction)

High frustration response: Hmm... Sounds like a clever move to me.

10. Would you recommend the game? Why/why not?

Control response: Interesting.

Low frustration response: Interesting (+interpunction)

High frustration response: Hmm interesting. Sounds like you know what you are talking about.

11. Is there anything left you would like to say?

Control response: Thank you for all the feedback. Bye bye.

Low frustration response: Thank you for all the feedback (+interpunction). Bye bye.

High frustration response: Thank you for all the responses. Your honest feedback is going to help a lot! Bye bye.