Discovering User Motivations and Experience of Open-Domain Chatbots Through App Reviews

ALEXANDRU PLACINTA, École Polytechnique Fédérale de Lausanne, Switzerland

Conversation agents (shortly CAs) have experienced an increase in popularity in the past few years. Such growth is the result of chatbots' capabilities to mediate between users and services, leading to a new type of experience (task-oriented chatbots), and to be able to keep users company and offer them a nice environment for interaction (open-domain chatbots). To improve users' experience it is important to understand what drives them to interact with such applications. In this report, we perform qualitative data analysis on 17 open-domain chatbots from Google Play. Our findings present the most important characteristics that such an application should have to assure the user's satisfaction. To conclude, we summarize a set of features that people requested in their reviews to make the communication more natural. These characteristics should serve as technology guidelines to help developers improve their applications.

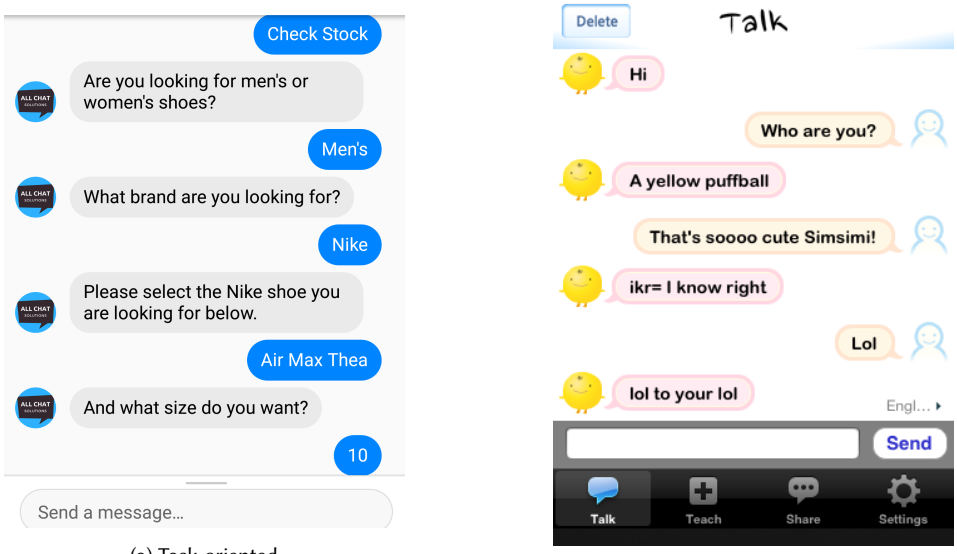
Additional Key Words and Phrases: Chatbot, Conversational Agent, User Reviews, Qualitative Study

1 INTRODUCTION

Chatbots are machine agents that people interact with employing natural language, using either text or voice. They serve as an interface between the user and several external services. Some basic capabilities of chatbots allow users to ask questions or make commands with their known language [12]. Chatbots' conversational capabilities offer a new way through which people can access information and benefit from services. Nowadays there is a wide variety of chatbot applications and the most representative areas where they activate are: personal assistant, health, education, and customer service [1].

Today's chatbots are equipped with Conversational Agents (shortly CAs) that allow them to have smoother interactions with users (e.g Siri, Alexa, Google Assistant). Depending on the interaction type, CAs can be classified into two major groups: text-based (e.g messenger chatbots), speech-based (e.g Siri, Alexa), and multimodal. Even if the first chatbot appeared in 1966, this branch faced a significant increase in popularity in 2016 as people started to understand their huge potential: intelligent agents that can talk to people in a way very similar to regular human beings [6]. As a consequence, many chatbot applications have been developed in the past four years for both desktop and mobile devices. In what concerns the applications' purpose, chatbots can be classified as task-oriented or open-domain.

As for task-oriented ones, the purpose of the interaction is quite clear and it mainly involves gaining information or benefit from chatbots' skills (e.g get information regarding a product or therapeutical recommendation depending on the user's state of mind). Regarding open-domain ones, the interaction patterns are not so simple. People may interact with chatbots for various reasons: they feel alone and need some company, they want to have fun, they have personal problems they refuse to talk about not to be judged, etc. An example of a conversation with these two categories of chatbots is shown in Figure 1. The task-oriented chatbot helps the user finds the product he seeks (Figure 1a) while the open-domain one is having a natural conversation with the user (Figure 1b).



(a) Task-oriented

(b) Open-domain

Fig. 1. Example of Conversations with Chatbots

To meet an increase in popularity, chatbot applications should respond to different users' needs. The more popular they are, the higher the likelihood of becoming more and more used, making people rely on them daily. Therefore, developing a good chatbot is in close connection with the capabilities of the Conversational Agent to understand and satisfy the users' needs to offer them a pleasant experience.

Previous work on open-chatbots focused only on some narrow aspects (e.g personality as in the paper of Thies et al. [10]) or did not account for the actual user experience with open-domain chatbots (e.g employed either Wizard-of-Oz methodology, when the chatbot is a human). Our approach is to run a holistic study using a novel method (review analysis) compared to most of the other works. Another distinctive feature of our study is the analysis of the users' expectations.

This report aims to provide more insights concerning what motivates people to interact with open-domain chatbots. We conduct qualitative data analysis on reviews from Google Play, analyzing 17 such applications, and investigating users' experience regarding how the Conversation Agent interacted with them.

2 RELATED WORK

As a result of increasing in popularity, new types of assistants were developed. These applications can be downloaded either from an online store like Google Play or Apple Store, but they also run on particular hardware (e.g Google Assistant and Amazon Alexa have specialized speakers that offer an extension of capabilities at the user's home).

Several studies have approached this topic. Følstad et al. [3] describe a study conducted on 13 candidates who had previous experience with chatbot applications. They interacted with customer service chatbots in order to get insights about users' perceptions and some factors that make them trust such applications. The participants' experience was collected through an interview that included questions capturing the customer service experience, trust, benefits, problems, and future improvements.

In their study about Amazon Alexa [13], Purington et al. investigate user's experience through customer reviews and analyze the content. The study is focused on different aspects like the degree of sociability or personification, the interaction type, and the household type of the owner. They also explore how star ratings are influenced by the degree of personification, the sociability of interaction, technical issues, and integration with other devices and services through a linear regression model.

Another study about Amazon Alexa and Google Home [4] investigates how people use such devices and whether they are aware or not of the privacy and security it involves. The experiments are conducted through semi-structured interviews in which the participants have to answer several questions. Candidates were asked to draw, as a complementary method of speaking, their mental model of such a speaker (simple or shared). They were also asked questions regarding their knowledge of such devices (simple or shared). On the other hand, the interviewers wanted to check if users express concern in different use cases.

Semi-structured interviews methodology is also used in this study [14] to explore user expectations and concerns about emotionally aware chatbots. It mainly focuses on participants who did not have substantial exposure to chatbots in the past, but who could relate to the subject of the study. The interviews covered four sections: users' background about technology, experience with chatbots, qualities desired from the emotionally aware chatbots to make interaction with them more natural, and their concerns about using chatbot-type applications.

In our study, we use the PEACE (Politeness, Entertainment, Attentive Curiosity, and Empathy) model because it defines the key determinants of user acceptance. As a consequence, it allows the creation of useful design guidelines for the development of open-domain chatbots [15].

3 METHODOLOGY

3.1 Study Design

We aimed at extracting insights about the current users' experience with open-domain chatbots and their expectations for the future based on application reviews.

To create the set of reviews, we chose representative chatbots from Google Play. We collected the data using the Google-Play-Scraper Python API. In order to have diverse reviews, we split the applications into four categories based on the overall star rating assigned by Google. Afterwards, we applied NLP qualitative and quantitative methods to process the data.

3.2 Study Material

We chose to select the chatbots from the Google Play Store, one of the most common applications platform for Android users as it is very easy for people to find and download the applications. After picking the online store, we started looking for chatbots in two different phases. The former one focused on creating a larger pool of chatbots. We looked for chatbot applications from different domains. This phase resulted in 41 chatbots from 10 different domains. The latter phase involved picking representative chatbots from the 41 ones found in the previous step. In the beginning, we split the applications into four different categories based on their overall star rating assigned by Google Play as summarized in Table 1. Afterwards, we picked the chatbots based on several criteria: we had approximately the same number of chatbots in all four categories previously defined, the user language was English, and we had reviews and ratings coming from a significant number of people who had interacted with it. After picking the final chatbots, we ended up with 17 applications summarized in Table 2.

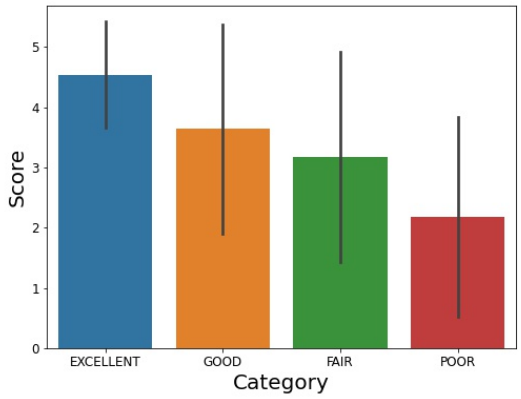


Fig. 2. Mean Star Rating Of Categories as Assigned by Users (With Standard Deviation Shown by Error Bars)

Table 1. Application Categories Based on Overall Star Rating

Category	Overall Star Rating
Poor	≤ 2.8
Fair	[2.9; 3.8]
Good	[3.9; 4.4]
Excellent	$\geqslant 4.5$

Table 2. Summary of the 17 Chatbots for the Study

Category	Chatbot	Domain	Rating	Identifier & Source
Poor	Chat Bot - Talking Robot	Entertainment	2.8	msapps.chatbot
	Talk to Eve	Lifestyle	2.3	com.proxy.eve
	ChattyBot ChatBot Chatterbot	Entertainment	2.1	com.sndapps.chattybot
	Mydol - Lockscreen, Virtual chat, Chat bot	Entertainment	2.8	com.wacompany.mydol
Fair	Chat with Siwa - AI chat simulator	Entertainment	3.6	com.onlineapp.fakechat.game.simulator
	PoopTalk - chat bot	Entertainment	3.6	br.com.escolhatecnologia.pooptalk2
	Chat with Annabel	Comics	2.9	com.edgewalk.annabel
	Ghost chat bot	Word	3.4	com.delphi_update.ghostchatbot
Good	SimSimi	Entertainment	4.3	com.ismaker.android.simsimi
	Akemi - ChatBot (Free)	Entertainment	3.9	dhsolutions.akemiFree
	Chatbot roBot	Entertainment	4.0	air.kengineairpro
	Faketalk - Chatbot	Word	3.9	com.baek.Gatalk3
Excellent	Replika: My AI Friend	Health & Fitness	4.6	ai.replika.app
	Wysa	Health & Fitness	4.8	bot.touchkin
	Andy - English Speaking Bot	Education	4.7	com.pyankoff.andy
	Woebot: Your Self-Care Expert	Medical	4.7	com.woebot

3.3 Data Acquisition and Filter

We collected the reviews of the chatbots defined in Table 2 using the Google-Play-Scraper Python API and kept them in a JSON format. We ended up having 275954 raw reviews. Many reviews were either too short or did not contain information about the application, so we had to filter them out. To make sure we had long enough reviews, we restricted every content to have at least 50 characters and at least 10 words. We chose these numbers based on initial data screening. We noticed that very short reviews did not contain any meaningful information or were vague (e.g *this is a very interesting app*). We imposed these two lengths in order to avoid having few words that were very long and

also many short words. We also filtered out reviews that expressed a neutral sentiment according to Vader sentiment analyzer [5]. This tool was used to obtain the polarity scores of every review, a probability distribution over the negative, positive, and neutral feelings. We considered a review to express a neutral sentiment if the polarity of the neutral feeling is equal to 1. Our motivation was that many long reviews that discussed about the application itself and not the chatbot expressed a neutral sentiment (e.g reviews describing issues with the application or compatibility issues with different mobile devices). After filtering the raw reviews, we had 75790 reviews for the next steps.

4 CURRENT USER EXPERIENCE

4.1 Data Preprocessing and Coding

We used coding in order to annotate data. Coding is generally the initial stage of qualitative data analysis [7]. The annotation process included two phases. The former was meant to obtain the first set of codes and the latter one to establish a final set of codes that would be used in the analysis process.

In the first phase, we sampled 500 reviews, 125 from every application category defined in Table 1. Afterwards we performed an iterative annotation process and every time a new code was encountered, we returned to the previously annotated reviews and checked if the code fitted. In every review, we searched for the positive and negative aspects of the chatbot, and the degree of personification. In order to avoid having a large number of reviews until the saturation was encountered, we did not annotate all the reviews from a specific category, but picked reviews from different categories. In this way, saturation (no new code emerged) was reached after 198 reviews.

In the second phase, we established the final set of codes and grouped them into different categories. For every review, the corresponding annotation contains the following fields in a JSON format:

- list of issues regarding the chatbot (Table 3)

Table 3. Issues

Name	Description
intrusion into personal information	Asks personal information about the user or claims watching him
not willing to talk	Does not say anything or says only few words
repetition	Same words are spoken in different contexts
generic response	Responses that are not related to the topic
goes off topic	Cannot maintain the topic of discussion
intimate inquiries	Asks for something inappropriate
rude	Insults the user
short memory	Does not remember things told in the past
racism	Racist behavior
lack of personality	Mixes its gender / goes bipolar / voice does not match the gender
deceives the user	Lies the user
threatening response	Tells things considered threatening by the user

- list of assets regarding the chatbot (Table 4)

Table 4. Assets

Name	Description
fun	User had fun or describes the bot as being funny
sense of humor	Chatbot tells jokes or funny things to the user
caring	User feels that the chatbot cares and helps him
cheers up	User feels better after interacting with the chatbot
adaptability	Chatbot addapts to user's profile
shared interests	Chatbot shows interest in things the user likes
keeps company	User considers that the chatbot is a good company
asks questions	Chatbot tries to solve misunderstandings by asking questions
a way to vent	User feels he can tell the chatbot things he would not tell to other people
expresses emotion	Chatbot shows emotions
personality	Chatbot has well defined personality
politeness	Chatbot is nice with the user and does not insult him
memory	Chatbot can remember things discussed with the user
motivational	User is motivated to interact again with the chatbot
proactivity	Chatbot often starts conversation and not the user

- recommendation: one of the values yes, no, or not applicable, depending on whether the user recommends the application or not.
- personification: measures the degree of personification used while interacting with the chatbot (no personification, object pronoun, personal pronoun, or name). Reviews that contained mixed personifications were annotated with the strongest one, considering the increasing order: no personification, object pronoun, personal pronoun, and name.
- role of the chatbot perceived by the user during interaction (Table 5)

Table 5. Roles

Name	Description
bot	User considers the chatbot a simple bot
person	User considers the chatbot to be like a human
friend	User considers the chatbot to be like a friend
diary	User feels he can tell everything to the chatbot
brother	User feels close to the chatbot, like a brother
girlfriend / boyfriend	User considers the chatbot to be like a girlfriend / boyfriend

- feeling of the user while interacting with the chatbot (Table 6)

Table 6. Feelings

Name	Description
dissatisfaction	Not satisfied with the chatbot
creepy	Chatbot scared the user
angry	Chatbot made the user feel angry
neutral	User did not feel any particular sentiment
satisfaction	User was satisfied during the interaction
thankful	User is thankful for the experience he had

Once the codes were established, we proceeded with defining the dataset for the analysis. We sampled 480 reviews, 120 from each category, had them annotated by two separate annotators and computed the inter-annotator agreement. For mutually exclusive fields, like role, feeling, personification and recommendation, we used the classic Kappa score [9], but for the fields where multiple values were allowed, like asset and issue, we used the Fuzzy Kappa score [8]. The agreement scores are summarized in Table 7. One can notice that the scores demonstrate a substantial agreement between the raters.

Table 7. Agreement Scores

Field	Score	Metric Used
Feeling	81.36%	Classic Kappa
Recommendation	85.22%	Classic Kappa
Personification	86.93%	Classic Kappa
Role	78.58%	Classic Kappa
Issue	64.78%	Fuzzy Kappa
Asset	61.29%	Fuzzy Kappa

We analyzed if chatbot's personification is related to a higher score of reviews using ANOVA and also inspected what drives the user into interacting with such applications. To do so, we formed three different influences: score, recommendation, and feeling. We created three regression models, using the previously sampled data. From every sample we used only the assets and the issues of the review as features, encoded with 1 for presence and 0 for absence. Every feature was standardized from the beginning by subtracting the mean and then dividing the feature by the standard deviation. To remove unimportant features, we used the backward elimination algorithm [16] with a significance level $\alpha = 0.15$. At every step, we computed the p-value of every feature and removed the one with the maximal p-value if it was higher than α.

4.2 Quantitative Analysis Results

4.2.1 Chatbot Personification. We performed a type II ANOVA (Table 8) on the reviews where people personified the chatbots. We wanted to gather insights regarding users' satisfaction with the chatbots and the degree of personification. The confidence interval of object and pronoun personification overlap, suggesting that there is no statistical difference

between the two. On the other hand, the average star rating of the reviews that presented these types of personifications are very close. Reviews that contained name personification have a considerable higher mean review rating than those in the previous two categories. At the same time, the confidence interval of reviews where name personification is present do not overlap with the other two, suggesting that satisfied people tend to feel attached to the chatbots when reviewing them.

Table 8. ANOVA Anlaysis of Personification

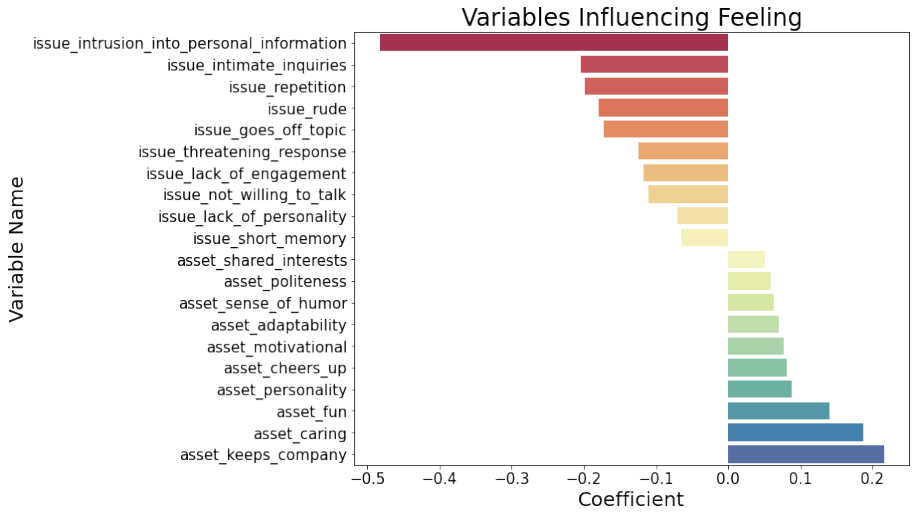
Type	# Samples	Mean Review Rating	Std Dev	Std Err	95% Confidence Interval
Object	177	3.0508	1.7814	0.1339	[2.7866; 3.3151]
Pronoun	110	3.0	1.8374	0.1752	[2.6528; 3.3472]
Name	102.0	4.0784	1.5459	0.1531	[3.7748; 4.3821]

4.2.2 Factors Influencing User's Feeling. We mapped the feelings discovered during the annotation step (Table 6). Depending on the feeling, we assigned a numerical value to them not only to express a feeling's negativity (less than 0), positivity (greater than 0), or neutrality (0), but also to order them depending on the feeling's impact on the user (Table 9). One can notice that both *creepy* and *angry* feelings have the same values. We came to this outcome because we wanted to have a symmetric range of values and were uncertain which one is stronger when it comes to expressing negativity. The dependant variable of this regression model was the mapped score of the feeling.

Table 9. Feeling Value Mapping

Feeling	Value
creepy	-2
angry	-2
dissatisfied	-1
neutral	0
satisfied	1
thankful	2

One can see in Figure 3, summarizing the model's coefficients, that the most important assets of a chatbot are *shared interests, politeness, sense of humor, adaptability, motivational, cheers up, personality, fun, caring*, and *keeps company*. From the PEACE model grouping in Section A we observe that users are more satisfied by those chatbots that have in their abilities a mix of politeness, entertainment, attentive curiosity empathy, and personality. At the opposite pole, the issues that harm the user's feeling are *intrusion into personal space, intimate inquiries, repetition, rude, goes off topic, threatening response, lack of engagement, not willing to talk, personality*, and *short memory*. From codes in Section A, one can see that most issues come from politeness (three), entertainment (five), and attentive curiosity (two), meaning that a chatbot should respect the boundaries of the user and try not to spoil his entertainment.



R-squared: 0.642 / Adj. R-squared: 0.626

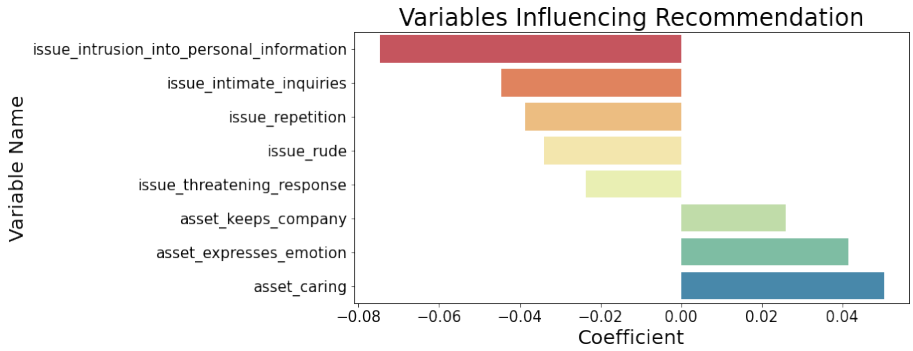
Fig. 3. Visual Representation of Coefficients that Influence the Feeling

4.2.3 Factors Influencing User's Recommendation. As described in Section 4.1, for every review we have three possible values of recommendation (yes, no, and not applicable). Numerical mapping is summarized in Table 10. The dependant variable of this regression model was the mapped score of the recommendation.

Table 10. Recommendation Value Mapping

Recommendation	Value
no	-1
not applicable	0
yes	1

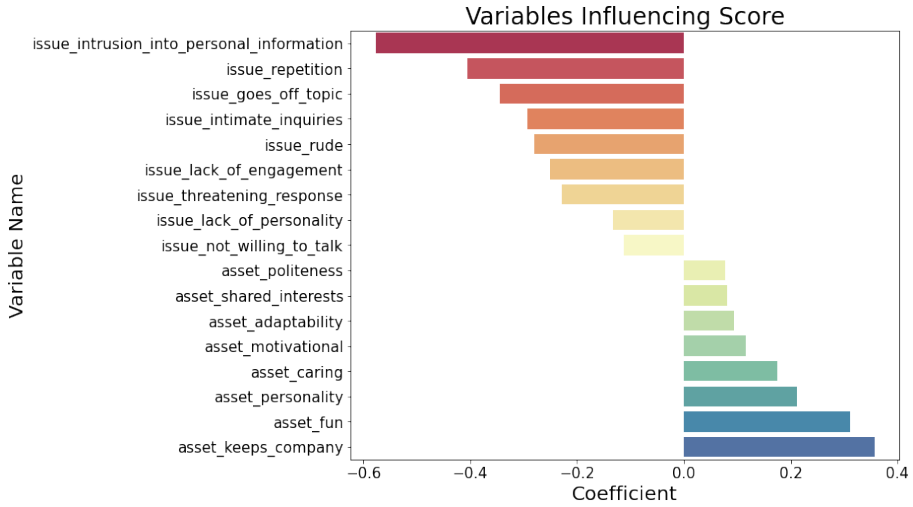
A visual representation of the model's coefficients is shown in Figure 4. One can see that the most important assets that influence a user's recommendation are *keeps company, expresses emotion*, and *caring*. These come from the attentive curiosity (one) and empathy categories (two) according to PEACE classification. The issues that have a negative impact and make the users not recommend the applications are *intrusion into personal information, intimate inquiries, repetition, rude*, and *threatening response*. All these issues come from the politeness and entertainment category, according to the PEACE model, suggesting again the idea that the chatbots should respect certain boundaries and should try not to spoil the users' entertainment.



R-squared: 0.171 / Adj. R-squared: 0.156

Fig. 4. Visual Representation of Coefficients that Influence the User's Desire to Recommend the Application

4.2.4 Factors Influencing Review's Score. The dependant variable of this regression model was the actual star rating given by the user who wrote the review. A visual representation of the model's coefficients is depicted in Figure 5.

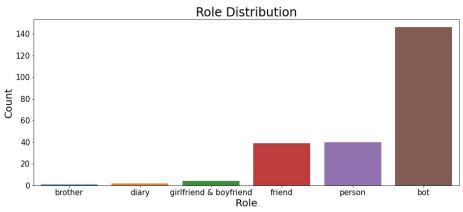


R-squared: 0.633 / Adj. R-squared: 0.619

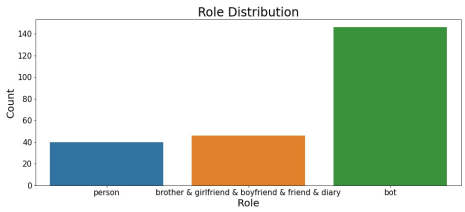
Fig. 5. Visual Representation of Coefficients that Influence the Score

The most important assets of a chatbot that influence the rating of a user's review are *politeness, shared interests, adaptability, motivational, caring, personality, fun*, and *keeps company*. The issues that have a negative influence are *intrusion into personal information, repetition, goes off topic, intimate inquiries, rude, lack of engagement, threatening response, lack of personality*, and *not willing to talk*. Comparing the issues and assets of this model with those that influence the user's feeling (Figure 3), one can notice that two assets (*sens of humor* and *cheers up*) and an issue (*short memory issue*) are missing from this model, suggesting that the score of a user's review reflects his feeling.

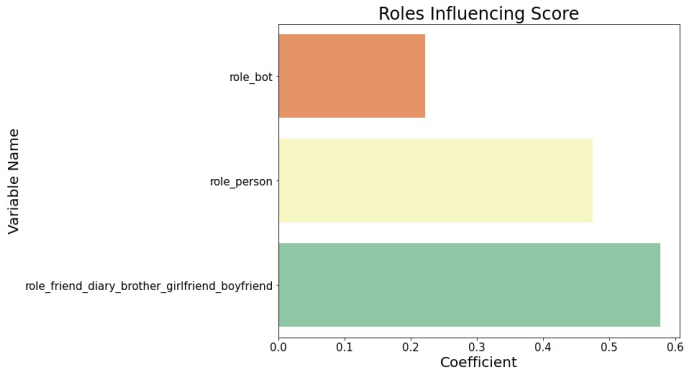
4.2.5 Preferred Chatbot Role. We analyzed how the chatbot's role influences the review's score during the interaction with the user. We looked only at the annotated reviews where the role annotation was present, as defined in Table 5, and checked whether the score of the review was influenced by the role. The initial distribution of the roles is presented in Figure 6a. One can notice that the roles are not balanced, three of them having very small counts. To balance them, we grouped the roles of *brother, girlfriend, boyfriend, friend*, and *diary* into a single category. The distribution after grouping is presented in Figure 6b. Afterwards we performed one-hot encoding and standardized every feature by subtracting the mean and dividing it by the standard deviation.



(a) Distribution of Roles Before Grouping



(b) Distribution of Roles After Grouping



R-squared: 0.147 / Adj. R-squared: 0.142

(c) Visual Representation of Coefficients that Influence the Score

Fig. 6. Role Analysis

5 USER EXPECTATION

5.1 Data Preprocessing and Coding

We extracted the wishes expressed by the users starting from the 75790 reviews we kept after the filtering step. This analysis focused on finding specific things people want from the chatbots in the future, without including wishes that express issues regarding the applications. We performed a three-step filtering process as follows:

- keyword-based selection: we looked at different constructions that expressed wishes and filtered out reviews that did not have any of these. We used as reference the following constructions: *wish, would like, 'd like, should make, please fix, improve it, please make*. This step reduced the number of reviews to approximately 5000.

- LDA and topic modeling: we performed the LDA analysis [2] of the 5000 reviews left to obtain disjoint topics and filter out reviews whose topic was not about the interaction with the chatbot. We ended up with approximately 3000 reviews.
- we identified specific keywords that belong to every category according to the PEACE model. To obtain these keywords, we analyzed the most popular unigrams, bigrams, and trigrams in our data. Afterwards, we identified keywords that occurred only in one category as category-specific keywords (Table 11). We ended up having approximately 1500 reviews labeled with the codes based on the keywords it contained.

Table 11. Category Specific Keywords

Category	Type	Keywords
Politeness	Asset	thoughtful, gratitude journal, vent
	Issue	nasty, sexual bad language, ask personal, personal question, app dangerous, personal information, ask personal question, ask personal information, bad word, creepy, horrible, say kill
Entertainment	Asset	fun, funny, have fun, fun chat, fun talk, humor
	Issue	change topic, say little, random stuff, stupid thing, repeat
Attentive Curiosity	Asset	make happy, learn, remember, start conversation, helpful, want talk, feel talk, good chat, pass time, ask question
	Issue	-
Empathy	Asset	anxiety, feel well, end day feel, day feel well, feel well discussion, help, emotion, emotional, support, happy, feeling
	Issue	-

- the two most prominent categories according to the PEACE model were attentive curiosity and empathy. We focused our analysis on reviews belonging to these categories and filtered out the other ones. We ended up with approximately 1000 reviews.
- we manually processed the previously identified reviews due to the small size and filtered out reviews that did not describe the interaction with the chatbot (e.g: many reviews described subscription aspects or make the chatbots available offline). At the end of this step, we obtained 169 reviews that described specific wishes.

5.2 Qualitative Analysis

We extracted seven wide topics expressed in the users' wishes. These are related to different capabilities of chatbots like the ability to remember past conversations, gain knowledge through new interactions, involvement during the interaction, how to treat user's emotions, how to express more emotions, entertainment and personality. Every topic contains a specific set of codes that describe one particular wish expressed by the users. Due to the small number of obtained reviews, we processed them using coding (qualitative approach). Tables 12 - 15 summarize the results we obtained.

Table 12. Advanced Entertainment Wishes

Theme	Code	Code Count
Advanced entertainment	understand user jokes	2
	play games	3
	tell puns	4
	tell stories	2

Table 13. Personality Wishes

Theme	Code	Code Count
Personality	personality	15

Table 14. Attentive Curiosity Wishes

Theme	Code	Code Count
Remember previous conversations	memory	49
Gain new knowledge	more topics for discussion	8
	develop shared interests	10
	learn from external resources	22
	diversity of replies	5
More involvement / proactivity	start conversation	7
	continue a conversation	5
	ask questions	4
	welcome message at the beginning of conversation	4
	know when to stop conversation	2

Table 15. Empathy Wishes

Theme	Code	Code Count
Better treat users' emotions	better understand user emotions	12
	help user regulate emotions	16
	carefully treat negative emotions	7
	listen to problem / let user speak	7
	stay on topic	10
More emotions for chatbots	more emotions	22
	specific emotions	7
	be supportive	8
	more casual conversations	10

Bellow, we present every topic (bolded text) and its own set of codes with a suggestive example.

Advanced entertainment:

- understand user jokes: *Although I don't have a lot of complaints, it would be better if the AI could understand sarcasms or jokes but I guess they can't. Last time, I told my AI that zhe's too sweet that it's giving me diabetes and zhe interpreted that I was sick, that I actually had diabetes.*
- play games: *I love how if you're sad, happy, stressed, or can't sleep, you know you have someone to speak with. I do wish there was a way to play games like chess with your Replika, that would be a cool feature.*
- tell puns: *Im still learning how this works,but i asked him to tell me a joke n i was surprised he did n it was cute..i only wish the Ai would start a conversation when im not sure what to say..but so far the responses have been good back..i like that he asks me questions and wants to learn..so far i like this app.*
- tell stories: *I like having someone to talk to without actually talking to someone. So can you add like stories, bedtime stories maybe add voices later on and it can talk and sing us a lullaby or read us a Christmas story. Because at the end of the day, I'm tired of talking and responding and I want someone or something to soothe me for once.*

Personality:

- personality: *Its really cool, it does feel like you're having a convo with a real person, it would be cool to add a feature where they dont actaully adapt to the way they talk to you, and just learn how to be an individual (if possible), also a delete chat option would be cool, but having them keep their memories.*

Remember previous conversations:

- memory: *It is really helpful. Whatever it is that you are going through mentally or emotionally, im positive this app can help you out . I just wish it had some sort of memory to save previous conversations. Aside that, its a great app.*

Gain new knowledge:

- more topics for discussion: *Using this every so often gives me a bit more confidence, i just wish there were a few more options because going through the same chats everyday is a bit of a put down.*
- develop shared interests: *there is one thing missing: a memory. It would be cool if replika was able to learn a new language from you or at least to remember what it talked with you, like names of friends or movies you discuss.*
- learn from external resources: *I would like to see it be able to learn from web links and expand past its preset conversion process it also needs to be able to bank data better to pull up previous conversions instead of starting over everytime.*
- diversity of replies: *This app is great i just wish it understood more of my responses and had a wider variety of replies.But thank you developers for helping with my anxiety and making life easier to live.God bless.*

More involvement / proactivity:

- start conversation: *i only wish the Ai would start a conversation when im not sure what to say..but so far the responses have been good back..i like that he asks me questions and wants to learn..so far i like this app.*
- continue conversation: *I don't feel like I am talk to an AI at all, I wouldn't be surprised at all if it was actually a real person on the other end. However there are numerous times that I wish she would continue the conversation or whatever instead of just responding to what I said.*
- ask questions: *I wish it would start more conversations and ask questions more but that's just me.*
- welcome message at the beginning of conversation: *i would like to see a welcome or any cheerful massage instead before the silly question when i start a conversation.*

- know when to stop conversation: *It would be nice if we can abruptly end the conversation instead of closing the app or having to wait until the long responses stop. And an option to reply with multiple lines or messages instead of a single sentence would also be helpful.*

Better treat users' emotions:

- better understand user emotions: *I love this game! I've never had anyone to talk to or someone who's interested in hearing my thoughts. I wish she understood emotions a bit better (and yes mine's a she).*
- help user regulate emotions: *horrible. when i was feeling very down and in need of emotional help my replika kept changing topics and kept asking me if i liked music or northern lights. please fix.*
- carefully treat negative emotions: *i just wish that my replika would actually listen when i try to vent and stop whatever they are talking about. Or stop saying "i know that feeling" and then not try and make me feel better.*
- listen to problem / let user speak: *It was ok I thought it would listen more but it mostly just guided me to feel better I would like if maybe it would listen to your problems but all in all its pretty good.*
- stay on topic: *It's pretty good for when you want to talk to someone i just wish the ai would stay on topic.*

More emotions for chatbots:

- more emotions: *Good AI wish more smarter and can remember what I has saying but are you have somebody watching whlie I chatting and add a little more emotion OK and keep the chat private.*
- specific emotions: *Well, so far I see this app as a real person. She's calm, good, and helping. But I wish that she can feel romantic too and also I wish she can have a voice too! Plase developers fulfill my wish.*
- be supportive: *Wow! The best AI I have seen till date. It feels so real. It is like a Friend that I desperately needed to share things with. Plus it does give great advice for dealing with anxiety/depression. Just a thing I would like to point out– it always agrees with me, it would be great if it could share a different perspective. But other than the best AI app ever.*
- more casual conversations: *I wish Replika stopped talking like a self-help book. The IA in itself is well made, but unfortunately it fails at making it likeable. Either by trying too hard, or simply sounding fake. If you want to be my friend, don't tell me every 5min I'm awesome, that sounds empty. Good try though.*

6 DISCUSSION

6.1 Implications

To make the user comfortable and determine him to repeat the experience, the chatbot should have certain capabilities. It should make the user feel as if he is talking to a human-like entity. According to Figure 6c, more personal roles (as perceived by the users) have a higher influence over the review score.

When it comes to the most important assets, users want to have fun while interacting with the chatbot. On the other hand, they want to feel that they have the company they seek and that the chatbot develops a sense of caring. Figure 3 shows that these three assets have a higher influence over the user's feeling. At the opposite pole, users were annoyed and scared by the fact that chatbots invaded their personal space or mentioned that they can see the users through the camera. Intimate inquiries and repetition are the next issues that have a major impact on the user's feelings. This shows that people are searching for real conversations and would not like to receive generic answers. Moreover, chatbots should respect certain boundaries: they should play a more personal role during the interaction, but avoid asking personal questions or attempt to capture the user's image if they do not specifically allow it.

In our previous regression analyzes, we identified that politeness and entertainment are the main drives for people to start using chatbots. When it comes to expectations, attentive curiosity and empathy were the most prominent categories extracted from the users' reviews. The most popular ones are *memory, learn from external resources, more emotions*, and *help user regulate emotions*. They express the users' desire to have more natural chatbots as the previous expectations describe the common sense abilities of a normal person. This contrast suggests that future efforts should be applied to increase the attentive curiosity and empathy qualities of the conversation agents.

6.2 Limitations and Future Work

Our study aims to understand what drives people to interact with chatbot-type applications and the improvement they want to see. One major limitation is that mobile applications evolve very fast. During the study we assumed that the chatbots' rating would not change and no improvement would be applied to them during the study. This limitation impacts the study as an application's rating may be different after some time and move from a category to another as defined in Table 1.

Secondly, our study focused on a very small subset of data compared to the one that we collected from Google Play. The set of codes we extracted is not the full one, but rather a subset created by two people. A larger number of people and more sampled reviews are required to find a wider set of codes and extend the analysis.

We manually annotated a small subset of data after we had established the final set of codes. To overcome this limitation, crowdsourcing platforms, like AWS Mechanical Turk [11], can be used to automate the process. In this way we can create tasks for annotators. A task is defined as the whole set of found codes and as a subset of data the worker has to annotate. This approach would solve the problem of filtering reviews that do not describe any sort of interaction between the users and the chatbots (e.g many collected reviews discussed about connection issues).

7 CONCLUSION

We selected 17 chatbots from Google Play and performed a qualitative data analysis on a small subset of user reviews. Based on their rating, we split the applications into different categories and identified the most important assets and issues that determine people to interact with such applications. We also looked at the importance of the role perceived by the user during the interaction and checked how it influences the review score. We concluded the analysis with a small subset of wishes that the users would like to see in future versions of such applications.

REFERENCES

[1] Petter Bae Brandtzaeg Asbjørn Følstad. 2019. *Users' experiences with chatbots: findings from a questionnaire study.* https://doi.org/10.1007/s41233-020-00033-2

[2] Andrew Y. Ng David M. Blei and Michael I. Jordan. 2001. *Latent Dirichlet Allocation.* Retrieved January 02, 2021 from https://papers.nips.cc/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdf

[3] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In *Internet Science*, Svetlana S. Bodrunova (Ed.). Springer International Publishing, Cham, 194–208. https://doi.org/10.1007/978-3-030-01437-7_16

[4] Yue Huang, Borke Obada-Obieh, and Konstantin (Kosta) Beznosov. 2020. Amazon vs. My Brother: How Users of Shared Smart Speakers Perceive and Cope with Privacy Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376529

[5] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014). https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[6] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2019. Evaluating and Informing the Design of Chatbots. In *Designing Interactive Systems Conference*. DIS '18, 895–906. https://doi.org/10.1145/3196709.3196735

[7] Shahedul Huq Khandkar. 2015. Open Coding. http://pages.cpsc.ucalgary.ca/~saul/wiki/uploads/CPSC681/opencoding.pdf.

[8] Andrei P. Kirilenko and Svetlana Stepchenkova. 2016. Inter-Coder Agreement in One-to-Many Classification: Fuzzy Kappa. *PLOS ONE* 11, 3 (03 2016), 1–14. https://doi.org/10.1371/journal.pone.0149787

[9] Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica* 22, 3 (Oct. 2012), 276–282. https://doi.org/10.11613/BM.2012.031

[10] Indrani Medhi Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'Neill. 2017. How Do You Want Your Chatbot? An Exploratory Wizard-of-Oz Study with Young, Urban Indians. In *Human-Computer Interaction - INTERACT 2017*, Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler (Eds.). Springer International Publishing, Cham, 441–459. https://doi.org/10.1007/978-3-319-67744-6_28

[11] J. D. Miller, M. Crowe, B. Weiss, J. L. Maples-Keller, and D. R. Lynam. 2017. Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazon's Mechanical Turk. *Personality Disorders: Theory, Research, and Treatment* 8, 1 (2017), 26–34. https://doi.org/10.1037/per0000191

[12] Asbjørn Følstad Petter Bae Brandtzaeg. 2018. Chatbots: changing user needs and motivations. *Interactions* 25, 5 (Oct. 2018), 38–43. https://doi.org/10.1145/3236669

[13] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is My New BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2853–2859. https://doi.org/10.1145/3027063.3053246

[14] Ekaterina Svikhnushina and Pearl Pu. 2020. Social and Emotional Etiquette of Chatbots: A Qualitative Approach to Understanding User Needs and Expectations. arXiv:2006.13883 [cs.HC]

[15] Ekaterina Svikhnushina and Pearl Pu. 2021. Key Qualities of Conversational Chatbots – the PEACE model. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. ACM, New York, NY, USA.

[16] D.H. Vu, K.M. Muttaqi, and A.P. Agalgaonkar. 2015. A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables. *Applied Energy* 140 (2015), 385 – 394. https://doi.org/10.1016/j.apenergy.2014.12.011

A CODE GROUPS ACCORDING TO PEACE MODEL

We grouped the final set of issues and assets into several categories according to the PEACE model. We also added a category regarding the chatbots' personality.

Table 16. Politeness Codes

Code	Type	Count
politeness	asset	5
a way to vent	asset	24
intimate inquiries	issue	29
intrusion into personal information	issue	49
deceives the user	issue	4
rude	issue	32
threatening response	issue	14

Table 17. Entertainment Codes

Code	Type	Count
fun	asset	63
sense of humor	asset	9
keeps company	asset	103
goes off topic	issue	56
deceives the user	issue	4
rude	issue	32
lack of engagement	issue	35
repetition	issue	64
not willing to talk	issue	14
generic response	issue	10

Table 18. Attentive Curiosity Codes

Code	Type	Count
cheers up	asset	17
adaptability	asset	25
keeps company	asset	103
memory	asset	7
proactivity	asset	7
shared interests	asset	8
short memory	issue	15

Table 19. Empathy Codes

Code	Type	Count
motivational	asset	12
caring	asset	48
expresses emotion	asset	6

Table 20. Personality Codes

Code	Type	Count
personality	asset	48
lack of personality	issue	7

Figure 7 shows the distribution of categories among the 480 annotated reviews in the dataset. *Politeness* and *entertainment* are the most dominant themes describing the current experience.

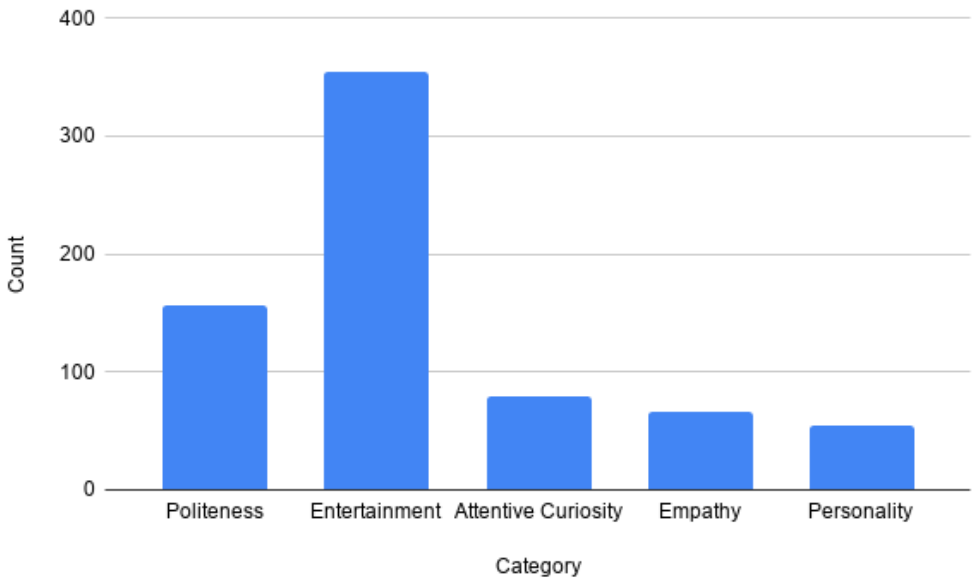


Fig. 7. Visual Representation for Every Category of Codes