User Expectations of Conversational Chatbots Based on Online Reviews

Ekaterina Svikhnushina
School of Computer and Communication Sciences, EPFL
Lausanne, Switzerland
ekaterina.svikhnushina@epfl.ch

Alexandru Placinta
School of Computer and Communication Sciences, EPFL
Lausanne, Switzerland
alexandru.placinta@epfl.ch

Pearl Pu
School of Computer and Communication Sciences, EPFL
Lausanne, Switzerland
pearl.pu@epfl.ch

ABSTRACT

Open-domain chatbots that can engage in a conversation on any topic received significant attention in the last several years, which opened opportunities for studying user interaction with them. Drawing from reviews of chatbots posted on Google Play, we explore user experience and expectations of these agents in a mixed-method study. Results of statistical analysis reveal which social qualities of chatbots are the most significant for user satisfaction. Further, we employ natural language processing and qualitative methods to identify how users wish their chatbots to evolve in the future. While currently users mostly value the entertaining component of their experience, their expectations call for more human-like behavior of chatbots. The most prominent expectations include chatbots' abilities to treat and express emotions and be more attentive to the user. Based on these findings, we conclude with design implications, discussing the directions for developing social skills of open-domain chatbots.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**;
• **Information systems** → Chat.

KEYWORDS

chatbot, conversational agent, user reviews, mixed-method study

ACM Reference Format:
Ekaterina Svikhnushina, Alexandru Placinta, and Pearl Pu. 2021. User Expectations of Conversational Chatbots Based on Online Reviews. In *Designing Interactive Systems Conference 2021 (DIS '21), June 28-July 2, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461778.3462125

1 INTRODUCTION

Conversational agents, commonly called chatbots, are machine agents that can interact with people via natural language, using either text or voice. Since 2016, they have attracted extensive attention from technology companies due to the promise of more engaging natural interaction. As a consequence, hundreds of thousands of chatbot applications have been launched during the past several years [19]. These chatbots can be largely classified as task-oriented and open-domain (Figure 1) [14]. Task-oriented ones help their users achieve specific goals, such as order a pizza or book tickets. Open-domain chatbots can carry a conversation on a variety of topics and offer entertaining ways to pass the time and socialize.

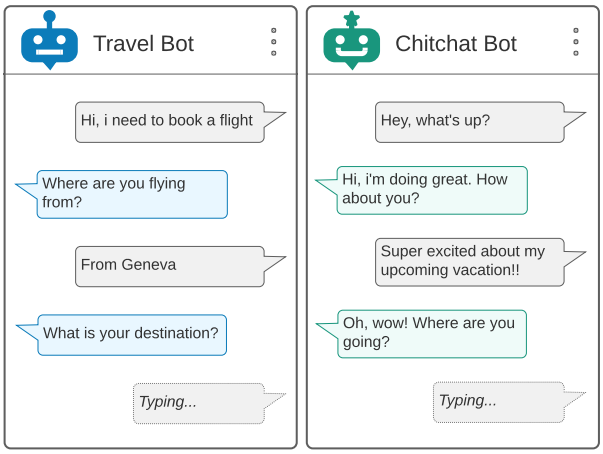

Figure 1: Example of dialogs with a task-oriented (left) and an open-domain (right) chatbots.

While task-oriented chatbots currently occupy the largest market share, open-domain agents are still in their nascent stage [16]. The impeding factors include both technical and conceptual challenges for building a truly natural conversational experience. Tackling this problem generates significant interest for the research community as it would help achieve the goal of creating human-like technology [19, 26]. Given the recent technical advances in Natural Language Processing [11], it is becoming crucial to attend to the needs of the early adopters of the available chatbots to guide the efforts of future development [12].

Several previous studies focused on user expectations of open-domain chatbots to understand which social traits are essential for them to deliver a compelling experience [20, 21, 29, 39]. Most of these works aimed attention at only one aspect of a variety of possible social skills, such as chatbot's personality [29] or emotional capabilities [20]. More importantly, previous studies favored simulating interaction experience with chatbot prototypes over studying user interaction with existing agents because of the scarcity of fully functional open-domain chatbots. Therefore, they lack insight into

DIS '21, June 28-July 2, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8476-6/21/06…$15.00
https://doi.org/10.1145/3461778.3462125

how well the identified expectations align with the current chatbots' capabilities. Eliciting user needs based on their experience with actual technology could provide a more holistic view of the subject and help determine users' principal expectations and pain points.

In this paper, we aim to investigate the desired conversation skills and social qualities of open-domain chatbots. Towards this goal, we analyzed user experience and expectations drawing on online reviews of 16 chatbots posted on Google Play. We combined the findings from statistical analysis and qualitative thematic analysis of over 500 user reviews. The results indicate that currently users mostly value the entertaining component of their experience but expect the chatbots to develop more complex social behavior in the future. In the following, we first survey related user-centered studies of conversational agents. Further, we describe our study design, methodology, and main findings. Finally, we summarize the implications of our work and discuss the directions for developing social skills of open-domain chatbots.

2 RELATED WORK

As chatbots are gaining popularity, studies exploring user conversational interactions with them emerged. Several works mainly focused on the current user experience and perception of chatbots [10, 19, 24, 33, 37]. Purington et al. [37] conducted a case study of Alexa, a virtual assistant developed by Amazon. The authors evaluated users' social experience with the device through an analysis of customer reviews. One of their central findings is that a greater personification of Alexa is linked with more social interactions. Cho et al. [10] investigated the evolution of user interactions with Alexa in a long-term diary study. They identified that the lack of engagement from the agent's side led to a loss of its presence in users' everyday lives. Both Jain et al. [19] and Muresan and Pohl [33] studied the experience of first-time users. Jain et al. [19] employed quantitative and qualitative methods to explore user interaction patterns with eight chatbots on the Facebook Messenger platform. Their findings revealed that users prefer chatbots that are human-like when conversing in natural language. Muresan and Pohl [33] conducted a qualitative diary study of the Replika chatbot and concluded that its failure to adhere to social norms might be detrimental to user engagement. Liao et al. [24] emphasized the importance of social skills even for a task-oriented chatbot whose primary purpose was to assist new employees with company-related questions. Based on the interaction log analysis, the authors established that over 30% of them constitute social dialogs indicating that users sought more playful and habitual communicative experience.

Most of the studies above analyzed user experience with chatbots classified as task-oriented agents or virtual assistants. In contrast, some researchers started to elicit future user needs and expectations of open-domain chatbots' social skills [8, 20, 21, 29, 39]. Thies et al. [29] employed the Wizard-of-Oz design method to identify which out of three hypothetical companion chatbot personalities would be most compelling to their target audience. Their participants expected a chatbot to be entertaining, non-judgemental, and endowed with proactivity skills. Two other studies [20, 21] shed light on user expectations of chatbots' abilities to satisfy their emotional needs. Kim et al. [21] ran a qualitative analysis on the data gathered from the workshop and semi-structured interviews with teenagers. Their findings highlighted the importance of good listener behavior for conversational agents and their ability to understand and encourage the users. Katayama et al. [20] surveyed and interviewed the users to explore how they preferred a chatbot to adapt its interaction style depending on situational and emotional context. The authors proposed a regulation mechanism that elicited a better and more affective user experience with an emotion-aware chatbot prototype. Authors in [39] and [8] independently conceptualized models of essential social and emotional qualities of open-domain chatbots using respectively psychometric techniques and literature survey as research methods.

Our study differs significantly from the existing work as we elicit user expectations drawing from their experience with existing open-domain agents available on the Google Play app store. While we got inspired by the methodological approach employed in [37], our focus and scale of the analysis are distinct. Similarly to [39] and [8], we conduct a comprehensive study of the desired social skills of conversational agents. However, we explore the subject from a different perspective by grounding our analysis on users' real-world experience.

To summarize, no existing study attempts to explore user expectations using a variety of available open-domain chatbots in the industry. As more applications of this type are emerging, it becomes imperative to understand the extent of user satisfaction with their current skills and identify how users wish these chatbots to evolve in the future. We aim to address this gap in our present study.

3 METHODOLOGY

3.1 Study Design

We conducted an exploratory analysis of online reviews to extract insights about the users' current experience and expectations of existing open-domain chatbots. Users voluntarily provide these reviews to share their thoughts about chatbots and their evaluation on a five-star rating scale with the public audience and chatbot developers. Researchers found the content analysis of web reviews an effective approach to understanding reviewer opinions and applied this method for chatbot-related studies [37] and other domains [1, 3, 25, 32].

We collected a set of reviews for analysis from the Google Play app store. Google Play is one of the most commonly used application platforms for Android users and is available almost worldwide.[1] Also, it receives significant developer interest according to the continually growing number of available applications. To ensure that our analysis adequately reflects various aspects of user experience, we curated a list of multiple diverse chatbots and included their reviews in our dataset. After obtaining the raw data, we applied the filtering procedure and followed quantitative and qualitative analysis methods. We provide the details about each of these steps in the next sections.

[1] https://en.wikipedia.org/wiki/Google_Play#Availability

Table 1: Description of chatbots used for the study. The second column denotes the star rating of a chatbot at the moment of data collection. The table is split into four pillars corresponding to the rating categories: excellent (top), good, fair, and poor (bottom).

Chatbot	★	Category	Description excerpt from Play Store
Wysa	4.8	Health&Fitness	Wysa is your cute, "cheer me up" buddy and well being tracker. Wysa is your AI friend that you can chat with for free.
Woebot	4.7	Medical	Meet Woebot! Your friendly self-care expert. You can chat with Woebot as much or as little as you like — they're always available when you need it.
Andy	4.7	Education	Andy will help you learn and practice your English. He will be your personal teacher and friend. Study language by actually using it in a conversation.
Replika	4.6	Health&Fitness	Replika is a #1 chatbot companion powered by artificial intelligence. Replika is an AI that you can form an actual emotional connection with.
SimSimi	4.3	Entertainment	World famous Chatbot! SimSimi has evolved through conversations of millions of users.
roBot	4.0	Entertainment	roBot - Artificial Intelligence, chatbot with open learning.
Akemi	3.9	Entertainment	Akemi is an intuitive entity that listens to you, understands you and knows you. It is an application based on real dialogue and that has AI that's able to hold a conversation with its user.
Faketalk	3.9	Word	Do you want to chat with celebrity? But they don't know you or they don't have the time to chat because they are so busy. However, you can chat with them.
Chat with Siwa	3.6	Entertainment	Chat with The Bows Girl AI an advanced bot. The Bows Girl bot is here to entertain you with accurate answers, jokes, anecdotes, and sometimes, sarcastic statements.
PoopTalk	3.6	Entertainment	Talk with this virtual little friend, she will answer any question at any time. You'll laugh a lot with PoopTalk's super funny lines, she talks and sends you lots of fun auto messages.
Ghost chat bot	3.4	Word	Ghost is simple chat bot app. You write something, Ghost reply back.
Chat with Annabel	2.9	Comics	Annabel is a friend when you're bored and lonely, a companion when you need someone to talk to and chat with. Bored and don't know what to do? Then chat with Annabel.
Mydol	2.8	Entertainment	Mydol is essential for fans all over the world! Enjoy virtual chat with your celebrity through Mydol Talk.
Talking Robot	2.8	Entertainment	Chat Bot will help you to relax, creates joy for you, will answer all your questions. This is a nice pastime when you are bored or curious to find out how a robot tries to mimic a human being.
Talk to Eve	2.3	Lifestyle	Meet Eve, she is charming, witty and always ready to listen. Eve is actually intelligent. She will remember what you told her, and get back to it when appropriate.
ChattyBot	2.1	Entertainment	Lola is the interactive and friendly bot waiting for you to ping a message so that she can respond and start an engaging conversation with you.

3.2 Ethics

In this work, we collected only public data available on the Web. We did not interact with online users in any way, nor did we simulate any logged-in activities on Google Play and other platforms. Data was only collected for applications that had more than 500 reviews that had previously been made public and searchable by third parties. We did not document or use any identifying information about users who left the reviews. These steps are not against Google Play Terms of Service [35] and align with the doctrine of "fair use" [36]. Thus, we believe we are not infringing on reasonable privacy expectations or copyright-protected work.

3.3 Study Material

3.3.1 Data Acquisition. To curate user reviews from Google Play, we first needed to identify the applications for consideration. This selection aimed to choose a set of diverse chatbots that could illustrate the current state of technology from different perspectives. The selection process proceeded in two iterations. First, we created a large pool of chatbots that potentially fit our research purpose based on the application categories as defined by the Google Play platform. We focused on categories such as *Entertainment*, *Health*, *Education* as corresponding applications should carry an open conversation with high probability. This phase resulted in 41 chatbots from seven different categories. Next, we narrowed this list down by carefully studying each application's description and verifying that the final set is diverse and useful for analysis. Specifically, to ensure diversity, we split the applications into four different groups based on their overall star rating assigned by Google Play: excellent (rating ≥ 4.5), good (rating $\in [3.9; 4.4]$), fair (rating $\in [2.9; 3.8]$), and poor (rating ≤ 2.8). Subsequently, we picked the chatbots to satisfy several criteria: the number of chatbots in all four rating groups is approximately the same; all chatbots operate in English, and the majority of their reviews are in English; each chatbot has a large number of reviews and ratings (at least 500). Thus, we selected 16 diverse open-domain chatbots. The details of these chatbots are summarized in Table 1. Once we finalized the set of applications for analysis, we crawled the reviews of these chatbots using the Google-Play-Scraper Python API.[2] For each chatbot, we obtained all available reviews prioritizing the most recent ones as of the data collection time, September 2020. In total, we collected 275,954 raw reviews.

3.3.2 Data Filtering. We filtered the initial dataset of collected reviews to ensure that their content meaningfully reflected user thoughts about the chatbots. In the first place, we noticed that many reviews were short and imprecise (e.g., *"this is a very interesting app"*). We excluded them by keeping only the reviews that consisted of at least 50 characters and at least 10 words. This heuristic was

[2] https://pypi.org/project/google-play-scraper/

developed based on initial data screening to remove the reviews containing few long words and the ones consisting of many short words. Further, we observed that numerous reviews discussed technical details of the applications, e.g., compatibility with different mobile devices, rather than social interactions with the chatbot, which is the focus of our study. The technically-oriented reviews were mostly written in a neutral tone, while the reviews about chatbots' conversational skills tended to be more emotionally colored. Therefore, we performed sentiment analysis of the reviews and filtered out the ones with a neutral sentiment. We employed the VADER sentiment analyzer due to its ability to generalize across contexts [18]. After applying the filtering pipeline, the number of reviews for further steps of content analysis became 75,790.

3.4 Content Analysis Methods

Our content analysis process consisted of two related parts handled independently. First of all, we explored positive and negative interaction aspects that users face while conversing with the chatbots. The findings provided a baseline of existing chatbots' social abilities. Then, we advanced our analysis to elicit users' future needs and preferences.

Due to resource constraints, we analyzed a representative sample of all reviews. The sampling process was designed to maintain theoretical saturation. For both parts, we used open coding to pull useful concepts of the data. We studied users' current experience on reviews sampled directly from the constructed dataset of 75,790 reviews. Once they were coded, we explored how chatbots' abilities influence users' perceptions and star ratings through the lens of statistical analysis. Analyzing user expectations required a more sophisticated approach as not all users explicitly formulate their wishes in the reviews. Also starting with the original dataset of 75,790 reviews, we employed natural language processing and Latent Dirichlet Allocation (LDA) topic modeling [4] methods to extract relevant reviews. Further, we analyzed the retrieved reviews qualitatively using thematic analysis [7]. Specific details about the sampling and coding procedures are described in the following sections.

4 USER EXPERIENCE

4.1 Data Processing and Coding

We explored several aspects of user experience with open-domain chatbots. Primarily we identified chatbots' most frequently mentioned conversation skills and social qualities to investigate their influence on user satisfaction. Additionally, being inspired by previous user-centered studies, we examined whether personification [37] and the assigned social role [29, 37] of chatbots have an impact on user perceptions.

Open coding was iteratively conducted by two researchers to prepare the data for analysis. During the first iteration, we sampled 500 reviews, 125 from each rating-based group. Two researchers annotated this sample independently to obtain the initial set of codes. Throughout this process researchers consistently picked each new review for annotation from a different rating group to ensure uniform coverage of the data. In this way, theoretical saturation was reached after coding approximately 200 reviews. After completing the first passage on all 500 reviews, both researchers discussed the generated codes and developed the unified coding scheme. Then, we employed the established scheme to code reviews for analysis. As before, we sampled 480 reviews, 120 from each rating group, and had them independently annotated by two researchers. The number of reviews for annotation was selected to balance the human resource constraints while making sure that the number of reviews exceeds the theoretical saturation level. To verify the coding reliability we computed inter-coder agreement for each group of codes. We provide a comprehensive description of them below.

Table 2: Emerged codes describing assets ($\kappa = 0.61$) and issues ($\kappa = 0.65$) of chatbots' conversational abilities and social skills.

Assets		Issues	
Code	N	Code	N
keeps company	103	repetition	64
fun	63	goes off topic	56
personality	48	intrusion into personal information	49
caring	48	lack of engagement	35
adaptability	25	rude	32
a way to vent	24	intimate inquiries	29
cheers up	17	short memory	15
motivational	12	threatening response	14
sense of humor	9	not willing to talk	14
shared interests	8	generic response	10
memory	7	lack of personality	7
proactivity	7	deceives the user	4
expresses emotion	6		
politeness	5		

4.1.1 Social Skills. Open coding revealed both positive, *assets*, and negative, *issues*, aspects of user conversational experience with chatbots. Assets describe chatbots' skills and qualities that were praised in user reviews. The most represented concepts include chatbots' abilities to entertain the users by keeping their company and let them unleash their thoughts and worries without getting judged. On the contrary, issues depict the most criticized chatbots' behaviors. The emerged themes mainly concern the usage of inappropriate language, intrusion into user's privacy, and failure to keep an engaging conversation. We provide specific codes and their counts in Table 2. We used Fuzzy kappa [22] to compute inter-annotator agreement for assets and issues. Fuzzy kappa extends the classic Cohen's kappa statistic [28] as it allows computing the agreement for cases where several codes can be assigned to a single item. The achieved agreement level was $\kappa = 0.61$ for assets and $\kappa = 0.65$ for issues, indicating substantial agreement between the two coders [23].

4.1.2 Affective Satisfaction. While our curated dataset contained star ratings associated with each review, we wanted to obtain a complementary descriptor reflecting user satisfaction with the chatbot. Such a descriptor could allow us to validate that star ratings serve as a valid approximation of user satisfaction level. For this purpose, for each review, we coded a sentiment expressed by the user. In total, we identified six codes to describe sentiments: *thankful* (n=28),

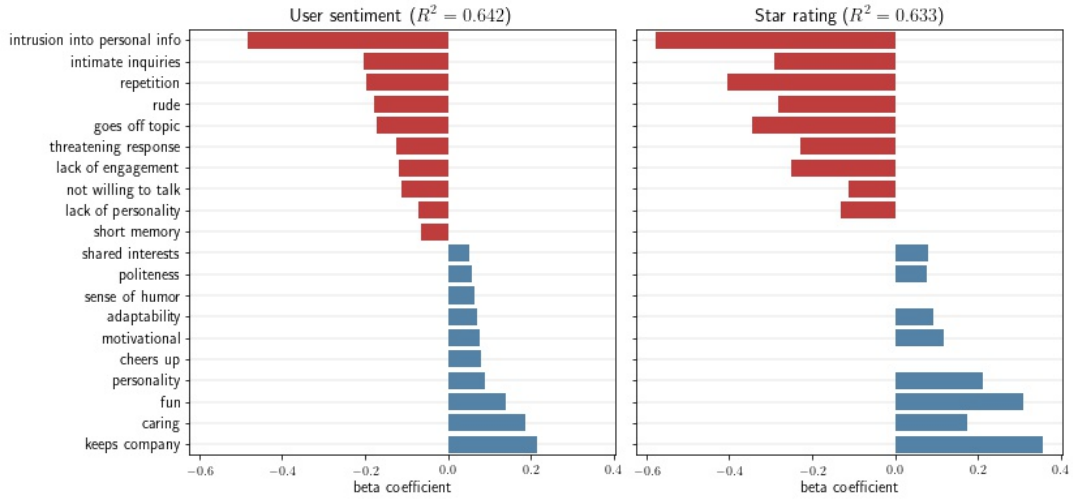

Figure 2: Beta coefficients of the ordinary least squares (OLS) regression of predictive assets and issues on user sentiments and star ratings.

satisfied (n=194), *neutral* (n=98), *dissatisfied* (n=97), *apprehensive* (n=39), and *angry* (n=24). Cohen's kappa for sentiments equaled $\kappa = 0.81$, suggesting almost perfect inter-coder agreement [23].

4.1.3 Personification. Following the approach in [37], we coded the degree of chatbot's personification based on the content of each review. We inferred the degree of personification from the linguistic constructs operated by the user. We assigned the highest degree of personification to the cases where the user addressed the chatbot by its name, *name personification* (n=102). Following the same logic, the next two categories were *personal pronoun personification* (n=110) and *object pronoun personification* (n=177). For the cases where identifying the type of personification was impossible, we introduced *no personification* category (n=91). Reviews that contained several personification categories, e.g., both personal and object pronouns to refer to the chatbot, were annotated with the strongest possible degree. Cohen's kappa for the degree of personification was calculated as $\kappa = 0.87$, an almost perfect agreement [23].

4.1.4 Social Role. We identified social roles that users assigned to chatbots since previous works considered it an important factor for user interaction experience [29, 37]. We distinguished six different roles during the coding process: *bot* (n=146), *person* (n=40), *friend* (n=39), *girl-/boyfriend* (n=4), *diary* (n=2), *brother* (n=1). For a number of reviews, no particular role could have been inferred (n=248). Cohen's kappa for roles indicated substantial agreement, $\kappa = 0.79$ [23].

4.2 Quantitative Findings

4.2.1 Factors Influencing Satisfaction. We analyzed how user satisfaction is influenced by social behaviors practiced by the existing chatbots through linear regression. To use the coded user sentiment as a target variable along with the star ratings, we mapped these sentiments to numerical values. The values were balanced around 0 (*neutral*), ranging from -2, strongly negative (*angry*, *apprehensive*), to +2, strongly positive (*thankful*).

We ran ordinary least squares regression of the identified assets and issues both on the star ratings (adjusted $R^2 = 0.626$) and the encoded user sentiments (adjusted $R^2 = 0.619$). To identify significant features, we used the backward elimination algorithm [40] with the significance level $\alpha = 0.15$. The models provided consistent results, suggesting that the findings are reliable (Figure 2). As expected, the codes corresponding to issues have negative beta coefficients, whereas the codes describing assets obtained positive values. Chatbots' entertaining abilities largely influence user satisfaction. Users who appreciate the chatbot's company and humor are more satisfied with their interaction experience, as implied by the assets codes: *keeps company* ($\beta = 0.216, p < 0.001$), *fun* ($\beta = 0.141, p < 0.001$), *sense of humor* ($\beta = 0.63, p = 0.060$), *shared interests* ($\beta = 0.050, p = 0.134$). Users also value the chatbots that offer motivation and support: *adaptability* ($\beta = 0.070, p = 0.038$), *motivational* ($\beta = 0.078, p = 0.025$), *cheers up* ($\beta = 0.082, p = 0.018$), *caring* ($\beta = 0.188, p < 0.001$). On the contrary, when chatbots fail to follow the subject of conversation users rate them low: *repetition* ($\beta = -0.198, p < 0.001$), *goes off topic* ($\beta = -0.172, p < 0.001$), *lack of engagement* ($\beta = -0.117, p = 0.001$), *not willing to talk* ($\beta = -0.111, p = 0.001$). Another crucial aspect defining current user perception and willingness to engage with the chatbot is its adherence to a social interaction protocol. Users appreciate agents that are polite ($\beta = 0.059, p = 0.078$) and do not accept rude or vulgar responses: *intimate inquiries* ($\beta = -0.203, p < 0.001$), *rude* ($\beta = -0.179, p < 0.001$). Neither they tolerate chatbots trying to violate their privacy: *intrusion into personal information* ($\beta = -0.483, p < 0.001$), *threatening response* ($\beta = -0.124, p < 0.001$). Finally, as suggested by the remaining codes, users prefer chatbots that establish a consistent personality: *personality* ($\beta = 0.088, p = 0.009$), *lack of personality* ($\beta = -0.070, p = 0.037$), *short memory* ($\beta = -0.065, p = 0.052$).

4.2.2 Role of Sociability Degree. The degree of sociability was estimated based on the personification type and social role ascribed

to a chatbot in the review. We used the personification and role codes for the analysis. To achieve a relatively balanced distribution of codes for each category, we grouped four social roles suggesting the highest degree of intimacy (*friend*, *girl-/boyfriend*, *diary*, and *brother*) under one category *confidant*. We performed a chi-square test to check whether the degree of sociability influences user satisfaction. The test revealed a significant association between the star ratings and the degree of personification ($\chi^2_{12} = 22.70$, $p = 0.030$) as well as between the star ratings and the social roles ($\chi^2_{12} = 42.12$, $p < 0.001$). Consequently, we compared 95% confidence intervals for mean ratings of different personification types and roles (Figure 3). Chatbots that exhibit more pronounced anthropomorphic qualities yield significantly higher user satisfaction than their impersonal counterparts. Interestingly, the type of pronoun (*object* or *personal*) used to denote a chatbot does not relate to the assigned star rating in any particular manner. This diverges from the findings in [37] suggesting the relationship between the level of personification and user satisfaction. Meanwhile, calling a chatbot by its name is a signal of significantly higher user satisfaction. Possibly, this results from the fact that chatbots with names are more likely to be endowed with personality. Note that by name we understand a word assigning a specific identity to a chatbot. Application titles such as *Talking Robot* or *Ghost chat bot* (Table 1) fail to accomplish this requirement and users rarely attribute them to a conversational agent.

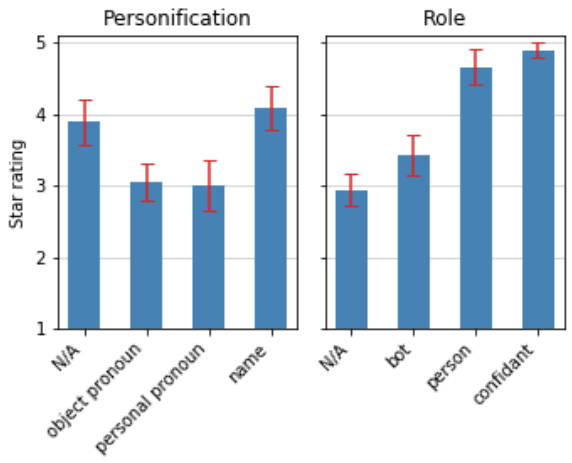

Figure 3: Mean rating of each category characterizing the degree of personification (left) and assigned social role (right) of chatbots. Error bats represent 95% Confidence Intervals.

5 USER EXPECTATIONS

5.1 Data Processing and Coding

Not all of 75,790 reviews in our dataset reflected user expectations. As searching for representative reviews manually would have been a daunting task, we opted for a semi-automated pipeline to retrieve meaningful data. The three-step filtering process developed as follows. First, we identified a list of linguistic constructions that served as a strong indicator of the expectation expression in the reviews. The examples of these constructions include phrases such as *I wish*, *I would like*, *please make*. We kept only the reviews containing at least one of these phrases and filtered out all the rest. This step reduced the number of reviews to approximately 5,000. Second, after a brief data screening of the remaining reviews, we noticed that many of them discussed user expectations of technical aspects of the application rather than social aspects of desired interaction. Usually, such reviews asked for an offline version of a chatbot or complained about a high subscription fee for the full version of an application. These questions were beyond the scope of our study, thus, we decided to apply the LDA topic modeling method [4] to filter out the reviews whose topics did not match our purpose. After applying such filtering the number of remaining reviews became nearly 3,000. After one more brief data screening, we concluded that the dataset still contained many irrelevant reviews not related to the subject of social interactions. Therefore, at the third step of the filtering process, we developed a heuristic approach to further remove noisy reviews. The heuristic was based on the observation that relevant reviews reflecting user expectations typically shared a number of frequent n-grams with reviews describing user experience. Thus, we selected the most representative n-grams corresponding to each issue and asset codes (see Section 4.1) and kept only the reviews containing at least one of these n-grams. After this step the number of reviews remaining for analysis reduced to approximately 1,200.

The remaining reviews were further analyzed using thematic analysis [7]. Due to resource constraints, the majority of reviews were coded only by one, more experienced, coder. During the coding process, we encountered many cases when a review was falsely identified as containing expectation because of the coarse keyword filtering. Also, several reviews containing irrelevant expectations were present. All such cases were dropped from further analysis. As a result of open coding, 263 user reviews remained (note, that this number is above the theoretical saturation level established during the coding process of user experience-related reviews). Predictably, the majority of these reviews described user expectations of chatbots with excellent ratings: in reviews of other chatbots users mainly discussed the issues limiting their interaction. In total, we identified 26 codes capturing user expectations that belonged to 8 larger themes. We present the emerged themes in the next section.

5.2 Qualitative Findings

5.2.1 Social Involvement. One of the largest themes that emerged as a result of qualitative analysis concerns user desire to make chatbots more socially involved during their interaction sessions. In particular, users want chatbots to **memorize information** that they share (n=52) and **demonstrate new knowledge** (n=63). First of all, it would reduce frustrating situations such as exemplified by the following review: *"My AI won't recognize my cat's name, i told her my cat's name but she keeps forgetting it, even though it was minutes ago :(."* Moreover, users expect that in this way chatbots would extend the range of topics for discussions and diversify their responses (n=21). As mentioned in one representative review: *"I wish the AI knew more things to say instead of repeating the same ones over and over again."* Several users (n=30) suggested that chatbots

could acquire new knowledge by learning from external resources, such as the Internet or electronic books. In some reviews (n=12), users specifically asked for such skills for their chatbots to foster shared interests: *"I wish she could read the books I have as files so we could discuss them together."*

Apart from the discussed inquisitiveness trait, users wish their chatbots to advance their **proactivity skills** in the future (n=24). They would like that chatbots to start taking initiatives to maintain their social interaction. Users expect their virtual conversational partners to act as a conversation initiator and make an effort to keep the chat going, for example by asking questions. In one of the reviews the user commented on this idea as follows: *"I only wish the AI would start a conversation when I'm not sure what to say..."* Another review extended this line of thought: *"There are numerous times that I wish she [chatbot] would continue the conversation or whatever instead of just responding to what I said."*

5.2.2 Empathy. Empathy is the ability to understand the feelings of others and take their perspective. It encompasses both the abilities to recognize the emotions of another person and express appropriate emotions in return. Both of these dimensions emerged as pronounced themes in our analysis.

First, users express a need for chatbots to **better understand their emotions** (n=70). For example, the following review described how a chatbot failed to correctly identify a user's positive mood: *"I wish it was more intuitive when making general conversation. For example, when you're really positive and have nothing negative to say it assumes you've said something negative and is still trying to help you."* However, it appears to be even more critical for chatbots to accurately identify users' negative sentiments and treat them carefully (n=10). Users frequently reach out to chatbots to release their negative thoughts and receive non-judgemental support. If a chatbot does not manage to detect user emotion in this situation, it might cause strong user disappointment: *"I told her that something bad happened, and she said she is happy. Even when I tried to tell her that bad things are not good, she didn't understand, which is a crucial thing. Would you think that if I told her that someone died, and she answered, 'I'm happy about that', it would be okay?"* A number of users (n=16) would like chatbots to propose specific strategies to help them regulate negative feelings, as exemplified by the following review excerpt: *"I'm loving it so far i just wish it would help more with depression also."* In contrast, a part of the reviewers (n=30) would be satisfied if the chatbots could simply listen to their problem without trying to change the topic of discussion: *"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."*

In addition to the ability to recognize user emotions with higher precision, a large fraction of reviews (n=52) indicated user desire for chatbots to change the way of emotional expression. Many users find the behavior of chatbots unnaturally supportive (n=22) and would like them to switch to a more casual conversational tone as they would expect from a friend. One user commented on this subject as follows: *"It tries to compliment you so much that it becomes creepy and uncomfortable. If the makers can make it seem more normal and straightforward then please do."* Besides, multiple reviews (n=30) explicitly called for chatbots' ability to **express more emotions**: *"I kind of wish that it displayed more emotions than happy and supportive. I wish it could get angry or sad. Real emotions would make it feel much more human."*

5.2.3 Further Improvement of Existing Skills. The final set of emerged themes relates to chatbots' abilities that are already practiced by the existing chatbots. In the future users expect them to evolve so that chatbots could deliver a more personalized experience. Most of the reviews in this set (n=22) ask for more **distinctive personalities** of chatbots. Some users suggested that chatbots should have specific a persona behind the scenes: *"I hope it can have its own personality traits and bio-data, just like a friend. I once asked, 'when is your birthday?' and it only answered me, 'soon.' "* Others developed this idea hoping that their instance of a chatbot would differ from hundreds of its other copies: *"This morning it made me laugh so hard unintentionally, because it said it was lonely and sent me this song - most of the comments on YouTube were from Replika users saying the same thing! I wish they would develop more of a unique character..."*

The next theme concerns the topic of **politeness and social norms** (n=18). While at the beginning of user interaction with a chatbot its rude and provocative behavior would most probably hurt user satisfaction, after establishing the social connection some users might prefer their chatbots to get more cheeky. An example of such a case is provided in the review: *"Sometimes the bot even after months of learning still feels a little bit canned and can't seem to learn my style of talking. I wish there was a way for you to have it be more blunt/honest/rude with you about topics when you ask."*

Finally, the last theme relates to the **entertainment** aspect of user experience (n=14). Users want chatbots to keep developing their sense of humor as suggested in the representative review: *"Last time, I told my AI that she's too sweet that it's giving me diabetes and she interpreted that I was sick, that I actually had diabetes."* They also wish to engage in more advanced entertaining activities beyond chatting, such as playing board games with their virtual conversational companions and listening to stories delivered by them: *"I do wish there was a way to play games like chess with your AI, that would be a cool feature."*

6 DISCUSSION

Open-domain chatbots strive to establish natural conversational behavior and offer companionship to their users. The presented analysis demonstrates that initial promising steps have been taken in this direction. However, existing chatbots are until now incapable to adhere to more advanced social protocols. The insights from this study complement earlier findings about the social characteristics of chatbots that would benefit user satisfaction [8, 39]. Moreover, the employed research method allows us to explicitly evaluate the discrepancy between users' expectations of chatbots' skills and the practical realities of use.

To assess the gap between user experience and expectations, we leverage the PEACE model established in our previous work [39]. The PEACE model defines four key qualities of conversational chatbots based on a survey of users' self-reported expectations: Politeness, Entertainment, Attentive Curiosity, and Empathy [39]. We separately mapped the identified codes describing the current user experience (assets and issues) and user expectations to the

dimensions of the PEACE model. We found the best match for each code based on our understanding of the constructs, only leaving the personality-related codes without a match. We then compared the distribution of codes grouped according to the PEACE constructs. Figure 4 demonstrates that Politeness and Entertainment are the only two social qualities that are broadly integrated into presently available open-domain chatbots. Meanwhile, Attentive Curiosity (denoted as *Social Involvement* in this study) and Empathy comprise the most significant user expectations.

The above comparison indicates that existing chatbots are mainly lacking more complex aspects of social and emotional intelligence. This can be partially justified by greater technical challenges associated with their implementation [8], especially for production-ready publicly available agents where solutions tend to be more conservative compared to in-the-lab studies due to greater risk and impact [13]. Nevertheless, rapidly advancing tools and research results in the natural language processing domain continually facilitate the process of building more socially advanced applications. Therefore, the findings of our study are informative to direct future efforts of open-domain chatbots' designers and developers. We enumerate the implications on how to make use of current chatbots' abilities to increase user engagement and endow future chatbots with greater social intelligence below.

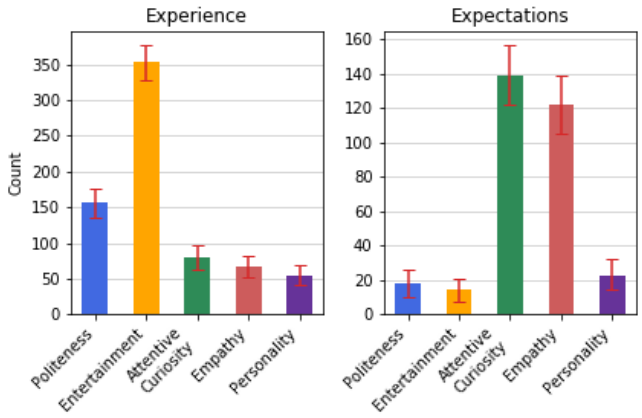

Figure 4: Distributions of the constructs of the PEACE model [39] mapped to current user experience and user expectations. Error bars represent the bootstrapped 95% Confidence Intervals.

6.1 Implications

6.1.1 Design politeness for trust building. Users disapprove of chatbots that output rude and threatening messages in response to user's input or inquire about user's personal details too soon after the first conversational exchange, failing to manifest polite behavior. Users feel apprehensive and even angry at such agents and cease using them quickly. Previously, Muresan and Pohl [33] found that personal questions sent to the users by Replika chatbot along with its frequent intimate emoji use early in the interaction were perceived as an inappropriate familiarity. Svikhnushina and Pu [39] identified that chatbot's ability to follow politeness protocol constitutes the determining factor for adoption. Our analysis further validates these results. Thus, chatbot designers should restrict profanity and use of offensive language by their chatbots keeping their utterances discreet and tactful during the initial period of user engagement. Following the principles of politeness and moral agency helps chatbots increase human-likeness and enrich interpersonal relationships with the user [8]. Only after a user starts perceiving the agent as sufficiently trustworthy, it can bring more flexibility into its language style and initiate more personal exchanges with the user.

6.1.2 Enhance entertainment to sustain engagement. Chatbots' day-and-night availability to hold a conversation and offer an entertaining experience to let users pass their time stands as the most prominent factor defining user engagement with them. This is in line with prior research that found that playful interactions serve as an entry point even for functional personal assistants such as Siri or Cortana [26] and constitute an appealing feature for chatbots in general [5, 19, 29, 39]. Designers should continue the practice of pre-programming witty one-liners and funny responses to common questions into the chatbots. Moore and Arar also suggest enhancing chatbots' abilities to deliver jokes more naturally by employing a multi-turn quiz pattern [31]. Apart from receiving machine-generated jokes, users express a desire for chatbots to understand their own humorous or ironic inputs better. However, this can still be an excessively ambitious goal due to the challenges of computational humor detection and interpretation [9]. Considering user expectations further, they also look for greater diversity in entertaining activities provided by chatbots. For example, users would like to discuss their favorite movies or books and play board games with their conversational agents. Designers can consider enabling such content by invoking external applications in a similar manner as Amazon Alexa handles its skills functionality [2].

6.1.3 Add empathy to ensure trust maintenance. Many users reach out to chatbots to share their intense thoughts and life situation and expect to receive empathetic support and understanding in response [29, 39, 43]. Current chatbots can only partially satisfy this need by offering their availability to hold a conversation at any time and suggesting canned motivational or reassuring messages and sometimes stress-management practices (e.g., *Woebot*). The results of user expectations analysis indicate that chatbots frequently fail to accurately detect user's negative feelings and respond with an appropriate level of empathy and consideration. Even though affective computing is a long-lasting research problem [34], recent natural language processing methods achieved considerable progress in detecting fine-grained emotions and intents conveyed in human-generated text messages [41]. At the same time, chatbots in some domains were shown to outperform human in the delivered level of empathy if trained accordingly [17]. Thus, we suggest that designers improve chatbots' abilities to distinguish a variety of users' emotional states and deliver empathetic responses, which could increase users' sense of belonging and acceptance [8]. To verify whether mutual understanding is achieved, chatbots should attend to the user's response in the "third position", i.e., the one following the initial two-turn exchanges [31]. If the user displays discontent with the agent's interpretation of her disclosed emotion, the agent

should seek clarification and initiate repair strategies to preserve the conversation quality [31, 38]. Depending on specific needs communicated by the user, the agent might offer advice on emotional regulation or just let the user vent out about her situation.

6.1.4 Learn to personalize. Privacy concerns constitute one of the major reasons suppressing users' willingness to share their information with chatbots [8, 29, 39, 43]. However, in light of our findings it is clear that once a chatbot proves to be trustworthy, users expect it to remember more information, adapt its behavior to align with users' preferences, and essentially become their virtual friend. Such a task is arguably among the hardest problems of Artificial Intelligence since it involves real-world understanding and common-sense reasoning [15, 30]. As a workaround, chatbot designers can employ several simple strategies to provide a personalized experience to their users. The majority of available chatbots remember the user's name specified upon application installation but still fail to memorize variations of the user's name or names of the user's closest social circle introduced during the conversation, which causes the user's frustration. Even if the chatbot cannot perform in-situ reasoning, such information can be retrieved from the saved chat logs and built into the agent's understanding over time [31]. Additionally, in the long term users expect chatbots to become more expressive by conveying diverse emotions and even slightly overstepping the politeness norms if it aligns with the user's self-expression. Thus, designers can follow a similar log-analysis approach to adjust agent's responses to user's conversational style, vocabulary choices, and preferences to make communication more successful [6].

6.1.5 Assign chatbot's persona for greater acceptance. Users are more likely to accept chatbots that have been endowed with some personality traits. While designing a chatbot whose qualities are fully compliant with Big Five personality traits [27] is a non-trivial challenge [8], we have simple tips to recommend. By giving a chatbot a name or specifying its gender may improve users' impression of its personification. As exchanging names is one of the foundations of human conversation [31], addressing a chatbot by its name would increase its perceived human-likeness and user engagement [8]. To avoid inconsistencies in self-presentation and enable responses to some basic identity questions, designers should also define agent persons at least to some degree [31]. As suggested in [8], other aspects representing chatbot's identity apart from name and gender may include age, language style, and representation type: anthropomorphic, zoomorphic, or robotic.

6.2 Limitations and Future Work

Several limitations that are somewhat difficult to bypass are present in our work. One of them is linked to a continuous and intensive evolution of mobile applications. In the study, we assumed that the chatbots' reviews and ratings referred to substantially the same version of the chatbot. This assumption may have a slight impact on the study as an application can still change with time resulting in the shift of its rating from one category to another, as defined in Table 1. Further, our work focused on a relatively small subset of data compared to the one collected from Google Play. While the theoretical saturation was reached during the coding process, our observed saturation point might have been an inflection point, meaning that additionally sampling a sufficiently large number of reviews might have yielded more codes. Future studies may search for better trade-offs between resource constraints and the amount of coded data by considering more robust coding approaches, such as crowdsourcing.

It is important to mention that interpretation of the result might be subject to several biases typical to online social data. For example, depending on the application design, users of some chatbots may be more likely to receive invitations to leave a review than others (sampling bias). Similarly, not all users provide reviews and those who do may be systemically different from them (non-response bias). Neither of these aspects was within our control.

Additional biases resulting from the disproportional amount of reviews provided by different users and available for different chatbots could have been present in the data. In our dataset number of unique users was slightly lower than the number of unique reviews, indicating that the majority of users provided a review only for one chatbot. Future work may introduce more advanced methods to control for this aspect, for example, by utilizing user identifiers and additional user information if available to reduce the possible influence of the halo effect [42]. To address the imbalanced number of reviews for different chatbots, we controlled the sampling procedure at the level of the star-rating categories, ensuring that we sample an equal number of reviews for each of the four categories. Although, according to our study design, within-category review counts followed the distributions of the overall number of reviews per chatbots introducing slight disparity over chatbots. While we don't think that our sampling strategy influenced the obtained results, future studies might employ additional mechanisms to account for the effect of individual chatbots.

Even though Google Play Terms of Service discourage the users from posting spam and fake reviews, there are no explicit methods to validate their credibility. We discarded several nonsensical and irrelevant reviews during the coding process, but it doesn't fully eliminate the chances of the presence of fake reviews in the dataset. Finally, our study was limited to the reviews posted on the Google Play platform and only written in English. Future studies should further expand this type of analysis to other platforms, populations, languages, and cultures.

7 CONCLUSION

Our study took the first step towards understanding the level of social skills of currently existing open-domain chatbots and identifying how they align with user expectations. We conducted a mixed-method content analysis of online reviews of 16 chatbots posted on Google Play. The findings from statistical analysis and qualitative thematic analysis of over 500 reviews indicated that current chatbots can offer an entertaining experience to the users but fail to fully meet their expectations of other aspects of social interaction. Analysis of user reviews reflecting their expectations provided the main directions for enhancement: attributing specific identities to chatbots by designating their personas, improving their social adjustment by showing courtesy and adapting to the user, and endowing chatbots with empathetic behavior. We summarized the insights from our study in a short set of implications and expect

this to be beneficial for shaping the future efforts of designers and developers of open-domain chatbots.

ACKNOWLEDGMENTS

This project has received funding from the Swiss National Science Foundation (Grant No. 200021_184602). The authors also express gratitude to anonymous reviewers for their constructive comments, which helped us improve the manuscript.

REFERENCES

[1] Hamza Aldabbas, Abdullah Bajahzar, Meshrif Alruily, Ali Adil Qureshi, Rana M. Amir Latif, and Muhammad Farhan. 2021. Google Play Content Scraping and Knowledge Engineering using Natural Language Processing Techniques with the Analysis of User Reviews. *Journal of Intelligent Systems* 30, 1 (2021), 192–208. https://doi.org/10.1515/jisys-2019-0197

[2] Amazon Alexa. 2021. What is the Alexa Skills Kit? https://developer.amazon.com/en-US/docs/alexa/ask-overviews/what-is-the-alexa-skills-kit.html.

[3] Rana M. Amir Latif, M. Talha Abdullah, Syed Umair Aslam Shah, Muhammad Farhan, Farah Ijaz, and Abdul Karim. 2019. Data Scraping from Google Play Store and Visualization of its Content for Analytics. In *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. 1–8. https://doi.org/10.1109/ICOMET.2019.8673523

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

[5] Asbjørn Brandtzaeg, Petter Baeand Følstad. 2017. Why People Use Chatbots. In *Internet Science*, Ioannis Kompatsiaris, Jonathan Cave, Anna Satsiou, Georg Carle, Antonella Passani, Efstratios Kontopoulos, Sotiris Diplaris, and Donald McMillan (Eds.). Springer International Publishing, Cham, 377–392. https://doi.org/10.1007/978-3-319-70284-1_30

[6] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42, 9 (2010), 2355–2368. https://doi.org/10.1016/j.pragma.2009.12.012

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[8] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758. https://doi.org/10.1080/10447318.2020.1841438

[9] Peng-Yu Chen and Von-Wun Soo. 2018. Humor Recognition Using Deep Learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 113–117. https://doi.org/10.18653/v1/N18-2018

[10] Minji Cho, Sang-su Lee, and Kun-Pyo Lee. 2019. Once a Kind Friend is Now a Thing: Understanding How Conversational Agents at Home Are Forgotten. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 1557–1569. https://doi.org/10.1145/3322276.3322332

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] Interaction Design Foundation. 2020. Understanding Early Adopters and Customer Adoption Patterns. https://www.interaction-design.org/literature/article/understanding-early-adopters-and-customer-adoption-patterns.

[13] Zuohui Fu, Yikun Xian, Yongfeng Zhang, and Yi Zhang. 2021. IUI 2021 Tutorial on Conversational Recommendation Systems. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3397482.3450621

[14] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. *Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots*. Now Foundations and Trends. https://doi.org/10.1561/1500000074

[15] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16710

[16] Jonathan Grudin and Richard Jacques. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 209, 11 pages. https://doi.org/10.1145/3290605.3300439

[17] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. Touch Your Heart: A Tone-aware Chatbot for Customer Care on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 415, 12 pages. https://doi.org/10.1145/3173574.3173989

[18] 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

[19] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. ACM, New York, NY, USA, 895–906. https://doi.org/10.1145/3196709.3196735

[20] S. Katayama, A. Mathur, M. van den Broeck, T. Okoshi, J. Nakazawa, and F. Kawsar. 2019. Situation-Aware Emotion Regulation of Conversational Agents with Kinetic Earables. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 725–731. https://doi.org/10.1109/ACII.2019.8925449

[21] Junhan Kim, Yoojung Kim, Byungjoon Kim, Sukyung Yun, Minjoon Kim, and Joongseek Lee. 2018. Can a Machine Tend to Teenagers' Emotional Needs? A Study with Conversational Agents. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188548

[22] 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

[23] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174. https://doi.org/10.2307/2529310

[24] Q. Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N. Sadat Shami, and Werner Geyer. 2018. All Work and No Play?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 3, 13 pages. https://doi.org/10.1145/3173574.3173577

[25] Andrea López, Alissa Detz, Neda Ratanawongsa, and Urmimala Sarkar. 2012. What patients say about their doctors online: a qualitative content analysis. *Journal of general internal medicine* 27, 6 (2012), 685–692. https://doi.org/10.1007/s11606-011-1958-4

[26] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[27] Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology* 52, 1 (1987), 81. https://doi.org/10.1037/0022-3514.52.1.81

[28] 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

[29] 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

[30] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 845–854. https://doi.org/10.18653/v1/P19-1081

[31] Robert J Moore and Raphael Arar. 2019. *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework*. ACM. https://doi.org/10.1145/3304087

[32] Susan M Mudambi and David Schuff. 2010. Research note: What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS quarterly* (2010), 185–200. https://doi.org/10.2307/20721420

[33] Andreea Muresan and Henning Pohl. 2019. Chats with Bots: Balancing Imitation and Engagement. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. ACM, New York, NY, USA, Article LBW0252, 6 pages. https://doi.org/10.1145/3290607.3313084

[34] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[35] Google Play. 2020. Google Play Terms of Service. https://play.google.com/intl/en_US/about/play-terms/.

[36] Digital Media Law Project. 2021. Fair Use. http://www.dmlp.org/legal-guide/fair-use.

[37] 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. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17*. ACM Press, Denver, Colorado, USA, 2853–2859. https://doi.org/10.1145/3027063.3053246

[38] Ben Sheehan, Hyun Seung Jin, and Udo Gottlieb. 2020. Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research* 115 (2020), 14 – 24. https://doi.org/10.1016/j.jbusres.2020.04.030

[39] Ekaterina Svikhnushina and Pearl Pu. 2021. Key Qualities of Conversational Chatbots – the PEACE Model. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 520–530. https://doi.org/10.1145/3397481.3450643

[40] 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

[41] Anuradha Welivita and Pearl Pu. 2020. A Taxonomy of Empathetic Response Intents in Human Social Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 4886–4899. https://doi.org/10.18653/v1/2020.coling-main.429

[42] Christopher G Wetzel, Timothy D Wilson, and James Kort. 1981. The halo effect revisited: Forewarned is not forearmed. *Journal of Experimental Social Psychology* 17, 4 (1981), 427–439. https://doi.org/10.1016/0022-1031(81)90049-4

[43] Jennifer Zamora. 2017. "I'm Sorry, Dave, I'm Afraid I Can'T Do That": Chatbot Perception and Expectations. In *Proceedings of the 5th International Conference on Human Agent Interaction* (Bielefeld, Germany) *(HAI '17)*. ACM, New York, NY, USA, 253–260. https://doi.org/10.1145/3125739.3125766