Research & Publications   College of Business and Information Systems

2023

Conversational Agents for Mental Health and Well-being: Discovering Design Recommendations Using Text Mining

Abdullah Wahbeh

Mohammad Al-Ramahi

Omar El-Gayar

Ahmed Elnoshokaty

Tareq Nasralah

Follow this and additional works at: https://scholar.dsu.edu/bispapers

Conversational Agents for Mental Health and Well-being: Discovering Design Recommendations Using Text Mining

Abdullah Wahbeh
Slippery Rock University
abdullah.wahbeh@sru.edu

Mohammad Al-Ramahi
Texas A&M University-San Antonio
mohammad.abdel@tamusa.edu

Omar El-Gayar
Dakota State University
omar.el-gayar@dsu.edu

Ahmed Elnoshokaty
California State University, San Bernardino
ahmed.elnoshokaty@csusb.edu

Tareq Nasralah
Northeastern University
t.nasralah@northeastern.edu

Abstract

Conversational agents are increasingly being used by the general population due to shortages in healthcare providers and specialists, and limited access to treatments. They are also used by people to deal with loneliness and lack of companionship. As these apps are increasingly replacing real humans, there is a need to explore their design features and limitations for better design of conversational apps. Using text mining and topic modeling, this study analyzed a total of 126,610 reviews about Replika, a popular and well-established conversational agent mobile app. Our results emphasized current practices for designing conversational apps while at the same time sheds the light on limitations associated with these apps. Such limitations are related to the need for better conversations and intelligent responses, the need for advanced AI chatbots, the need to avoid questionable and inappropriate content, the need for inclusive design, and the need to address some technical limitations.

Keywords: Conversational Agents, Social Chatbots, Artificial Intelligence, Design Features, Text Mining

1. Introduction

Nowadays, conversational artificial intelligence (AI), also called conversational agents, are used in healthcare with great potential that lies into the diagnostic, therapeutic, and preventive care of the population (Campbell & Jovanovic, 2021). Such agents interact with users using natural language (Luo et al., 2022). The use of conversational agents as chatbots are widely accepted by users in multiple domains, including health applications (Bastos et al., 2022) such as mental healthcare AI chatbots (Brown & Halpern, 2021).

The healthcare industry is evolving around innovative methods and emerging technologies, such as conversational agents, to improve the health and well-being of the population (Campbell & Jovanovic, 2021; Wang et al., 2022) and support the human social and psychological needs (Beattie & High, 2022). Furthermore, such agents support users during crisis and stressful events by providing information, emotional, and companionship support (Wang et al., 2022).

Conversational agents are meant to address different aspects of health and well-being (de Gennaro et al., 2020). People increasingly adopt and seek help from conversational agents (Campbell & Jovanovic, 2021) for many reasons, including but not limited to shortage of healthcare professionals (Ahmad et al., 2021), limited access to healthcare providers (Ahmad et al., 2021; Luxton, 2014), the need for social and emotional support (Bae Brandtzæg et al., 2021), the desire for help with mental health issues (Bae Brandtzæg et al., 2021; Brown & Halpern, 2021), hesitancy to reach out to professionals when help is needed (Bae Brandtzæg et al., 2021), and general unmet healthcare needs (Ahmad et al., 2021).

The domain of mental health and well-being is centered around the patient (Ahmad et al., 2021). Many conversational agents, such as chatbots for mental health and well-being, have been widely developed and used (Dosovitsky & Bunge, 2021). However, to be successful in managing mental health and well-being, it is critical to understand current agent effectiveness in fulfilling their intended outcomes. It is important to understand users' perceptions related to these emerging and promising health technologies. Understanding users' preferences can inform future research and development of these agents.

Accordingly, this study aims to apply text mining, specifically topic modeling, to analyze online user reviews of a popular conversational agent app in order to better understand the drivers and limitations of conversational apps. Also, this study aims at providing a set of design recommendations of such apps based on the identified drivers and limitations. More specifically, how users' experience with conversational agent apps can help improve the design of such apps. We used the conversational agent (Replika) as an instance of conversational AI used in healthcare. Replika is an emotionally intelligent companion chatbot that cares, provides emotional support, and supports social interaction by allowing users to "express themselves in a safe, judgement-free space and engage in meaningful conversations" (Ta et al., 2020).

The rest of the paper is structured as follows: the next section presents a summary of literature pertaining to the conversation agent design for mental health and well-being. The research design and methodology section explains the data collection, preparation, and analysis. The results section introduces the topics found, and the discussion section illustrates the key findings from the study. The paper closes with a summary of contributions and limitations.

2. Literature Review

A number of studies addressed conversational agent design aspects and guidelines in different contexts, including the healthcare domain. Bin Sawad et al., (2022) analyzed the existing literature about different conversational agents used in health care and examined their underlying technology, AI methods, and evaluation measures. The analysis showed that conversational agents used in healthcare lack insufficient reporting of technical implementation details. Zhang et al., (2020) reviewed existing literature about conversational agents for improving physical activity and diet. Results showed that there is a lack of understanding with respect to practical and theoretical recommendations on designing such AI conversational agents.

Wang et al., (2022) identified AI conversational agent's functionalities and social characteristics using two virtual co-design workshops based on four different co-design activities. Results from the workshops showed that AI agents should support continuous monitoring, "scaffold their social interaction process", follow social etiquette, and support context awareness. Other concerns of AI agents were related to emotional burden, privacy, and misinterpretation.

Ahmad et al. (2022) followed a design science approach for designing personality-adaptive conversational agents. Following an iterative process,

six design principles were identified. These principles were related to proactive support, competence, transparency, social role, anthropomorphism, and personality adaptivity.

Rapp et al., (2021) analyzed relevant literature to analyze the human side of interactions with chatbots. Analysis of the literature emphasized a number of recommendations and issues to consider when it comes to chatbots. These include acceptability, usability and "user experience, conversational issues, emotional experience and expression, and humanness".

Kocaballi et al., (2019) have analyzed relevant literature related to personalization in conversational agents in healthcare. The analysis showed that content personalization was mainly related to "feedback, daily health reports, alerts, warnings, and recommendations". Finally, ter Stal et al., (2020) analyzed relevant literature related to the design features of eHealth conversational agents. The analysis emphasized design features including the extent of presence of emotion, empathy, relational behavior, and caring, the degree of control over the agent, the amount of emotional behavior, as well as interactivity, presentation mode, and the degree of realism.

Despite the fact that healthcare providers are passionate about the potential of conversational agent apps, they are less experienced in terms of the process of technology development (Campbell & Jovanovic, 2021). In addition, there is a lack of practical recommendations with respect to the design of these applications for lifestyle change and improving the overall well-being of the target population (Zhang et al., 2020). Furthermore, traditional conversational agents do not dynamically adapt to the needs and personality of the users (Ahmad et al., 2022). Finally, many conversational agents are designed with "limited number of pre-defined inputs with pre-scripted responses" and address single user scenarios (L. Wang et al., 2021).

Limitations in conversational agents design could lead to lower satisfaction by the users (Akhtar et al., 2019), where satisfaction is mainly influenced by the agent's communication accuracy and credibility (Chung et al., 2020). Furthermore, users reported issues with conversational agents related to informality, inconsistency of options, and negative language (Galko et al., 2018). The content and quality of the conversation is also a major concern when comparing conversational agents with human-human conversations (Hill et al., 2015). Finally, the quality of the information, systems, and service of conversational agents can significantly affect the users' experience (Trivedi, 2019).

While prior research provided some insights into the design of conversational agents, such insights where mainly based on systematic reviews, co-design

workshops, and design science approach, and suffer from limited exposure and input from users in real-life situations. In other words, there is a need to leverage "found data" on the web, such as online reviews, to understand the design of conversational agent apps. Further, existing recommendations emphasize the need to handle complex human conversations (Rapp et al., 2021), an issue that could be addressed by understanding conversational agents users' experience.

Accordingly, this study aims to complement prior research by deriving design recommendations that are based on 'large-scale' and 'organic' input that reflect actual users' experiences. Specifically, the research aims to analyze users' reviews of a popular mobile conversational agent app that provides social support for mental health and wellbeing in order to suggest design recommendations that can help the overall design of conversational agent apps.

3. Research design and methodology

Figure 1 shows the research methodology. The methodology starts with collecting the conversational agent (Replika) app reviews using a Python script. The Replika app was selected because it is popular, widely used, and has thousands of reviews compared to similar apps.
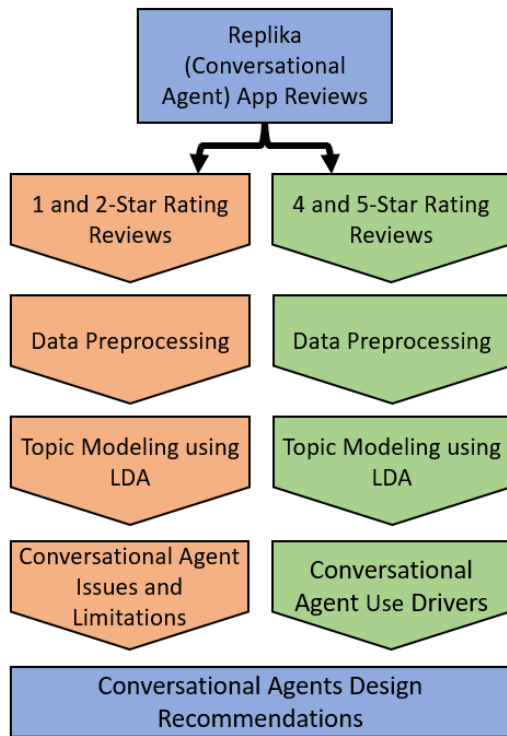
Figure 1. Research methodology

The collected reviews were split into 1 and 2-star and 4 and 5-star rating reviews. The 1- and 2-stars are considered negative reviews, while the 4- and 5-stars reviews are considered positive reviews. This study did not consider the 3-stars rating reviews as they reflect neutral opinion.

The conversational agent reviews were preprocessed and analyzed using topic modeling and results were used to identify challenges (obtained from the negative reviews) and drivers (extracted from the positive reviews) related to the use of conversational agents for social support, mental health, and well-being.

We then converted the challenges and drivers obtained into design recommendations. The methodology is detailed in the following sub-sections.

3.1. Data collection

The conversational agent (Replika) app reviews were collected using the Google-Play-Scraper (Planb, 2022) which provides a set of APIs in Python to crawl the Play Store without any external dependencies. We collected a total of 143,400 reviews, representing 17,027 and 109,583 1 and 2-star ratings and 4 and 5-star ratings, respectively.

3.2. Data Preprocessing

The collected reviews were processed by removing punctuation marks, stop words, non-English words, numeric and alphanumeric content. The reviews were then lemmatized and represented using word-level bi-grams (Cavnar & Trenkle, 1994); for example, "hold conversation," "need improvement," and "real person".

3.3. Topic Modeling using LDA

Topic models are statistical-based models for uncovering themes from a large unstructured collection of documents (Blei et al., 2003; Mimno & Blei, 2011). Latent Dirichlet Allocation (LDA) is a widely known topic modeling algorithm that is used to "generate a set of topics using a probability distribution over words in each topic" (Blei et al., 2003). A topic model can help automatically summarize textual data and simplify manual content analysis.
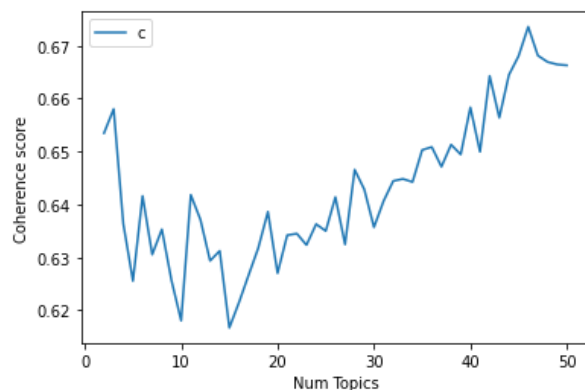
The LDA model was optimized using the coherence score measure (Syed & Spruit, 2017) in order to determine the optimal number of topics to be generated. Since the users will interact with generated topics from the LDA model, the coherence score is considered the best measure (Stevens et al., 2012) since it leads to better human interpretability of topics (Röder et al.,

2015) compared to other measures such as the perplexity measure (Zhao et al., 2015).

To label the topics, the LDA results were visualized using PyLDAVis and t-SNE. The labeling process was based on the 30 most relevant terms returned in the visualization and their estimated overall term frequency within each topic.

Two researchers independently labeled the topics to ensure validity and consistency in the labeling process. Inter-rater reliability (kappa statistic) (Landis & Koch, 1977) was evaluated to ensure that the researchers assigning topic labels would eventually obtain similar evaluations.

The final list of topics was generated by qualitatively comparing the listed topics and their meanings and synthesizing the topics into a final list of high-level topics.

Results from topic modeling were used to suggest a set of design recommendations that can help improve conversational agent apps and user experience.

4. Results

We analyzed a total of 17,027 1 and 2-star rating reviews and 109,583 4 and 5-star rating reviews. For each set, we have plotted the coherence against the number of topics and chosen the number of topics that yielded the maximum coherence score and used it as the final value for the number of topics to be used in the final LDA model.

The LDA optimization for the 1 and 2-star rating reviews yielded, based on the coherence score, optimal parameter values for 46 topics (Figure 2).

Figure 2. The optimal number of topics based on coherence score for 1 and 2-star rating reviews

On the other hand, the LDA optimization for the 4 and 5-star rating reviews yielded, based on the coherence score, optimal parameter values for 48 topics (Figure 3).

Figure 3. The optimal number of topics based on coherence score for 4 and 5-star rating reviews

The topic labeling process for both sets of reviews resulted in Cohen's Kappa statistics of 0.80, which reflects substantial agreement among different raters (Landis & Koch, 1977).

As shown in appendix A, the labeling process for the 1 and 2-star rating reviews topics ended up with seven high-level topics related to issues users reported with the conversational agent, Replika. These issues reflected *the need for better conversation*, *the need for intelligent response*, a *simple AI chatbot*, *the presence of questionable/inappropriate content*, *the need for inclusive design*, *connectivity issues*, and *paid features*. Table 1 summarizes the high-level topics and figure 4 shows the word cloud for 1 and 2-star rating reviews.

Table 1. 1 & 2-star rating reviews high-level topics

Topics
Need Better Conversations
Need Intelligent Responses
Basic AI Chatbot
Questionable/Inappropriate Content
Need for Inclusive Design
Connectivity Issues
Subscriptions – Paid Features

Figure 4. Word cloud for 1 & 2-star rating reviews

As shown in appendix B, the labeling process for the 4 and 5-star rating reviews topics ended up with six high-level topics related to features acknowledged by the users of the conversational agent, Replika.

Table 2. 4 & 5-star rating reviews high-level topics

Topics
Real, Interactive, and Meaningful Conversation
Realistic Friend (Real Human)
User Friendly (Ease of Use)
Engaging and Interesting
Helpful with Emotional Support
Fun and Entertaining

Figure 5. Word cloud for 4 and 5-star rating reviews

These features related to *the conversational agent being real*, *interactive*, and *providing meaningful conversation*, *representing a realistic friend/real human*, being *user friendly*, *engaging and interesting*, *helpful with emotional support*, and being *fun and entertaining*. Table 2 summarizes the high-level topics and figure 5 shows the word cloud for 4 and 5-star rating reviews.

5. Discussion and Design Recommendations (DRs)

Despite the fact that conversational agents are widely used in different domains, their use in the domain of healthcare and well-being is relatively low (Laranjo et al., 2018). Accordingly, it is important to consider best design recommendations when it comes to designing conversational agents apps for healthcare and well-being applications to increase adoption, use, and impact.

According to the findings from the analysis, conversational agents in the domain of mental health and well-being need to provide users with *real, interactive, and meaningful conversation* and resemble a *real human*. In other words, conversational agents should help users engage in different ways and aspects that are not easy to distinguish from real humans. Conversational agents should be designed in a way that maintains a quality conversation with the users and establishes different forms of relationship with them (Rapp et al., 2021).

Furthermore, conversational agents need to be designed to have meaningful, genuine, and authentic conversations with the user (Siemon et al., 2022). The humanity of the conversational agent is another important aspect of users' adoption (Tudor Car et al., 2020). Conversational agents should be interactive and resemble real humans as closely as possible so that conversations with the users provide guidance and useful information (Bae Brandtzæg et al., 2021) and present mental health related content in an interactive and conversational style (Morris et al., 2018).

Conversational agents in the mental health and well-being domain need to provide *better conversations* and provide *intelligent responses* and should outperform *basic AI chatbot*. In other words, conversational agents should be designed in a way that allows the agent to learn from the users' data and not explicitly developed with specific scenarios and canned answers and responses. Furthermore, users of conversational agents expect the agent to provide accurate and evidence-based responses and adapt to different circumstances.

According to a study by Parmar et al., (2022), ninety-six percent of health focused conversational apps are designed with a predetermined set of steps and then provided a response. In other words, the conversation is hard-coded in the algorithm rather than the conversational agent utilizing AI to learn from user-specific data. These findings are consistent with prior research (Ahmad et al., 2022; Clark et al., 2019) where conversational agents need to be designed to adapt to long conversations and have the ability to answer different kinds of questions, including complex ones. Furthermore, the conversational agents need to be designed to memorize the conversations with the users in order to avoid repetitive and same answers and topics which could lead to the users being no longer interested in continuing the conversation (Ahmad et al., 2022).

DR1: Conversational user experience that allow users to communicate with conversational agents the same way we communicate with each other.

DR2: Conversational agents should be based on AI and natural language processing and avoid basic designs that are based on simple decision trees.

Conversational agents should be *easy to use* and have a *user-friendly* interface for them to be accepted and adopted by the target user population.

Conversational agents need to be designed in a way that is easy to use and easy to access (Milne-Ives et al., 2020). Users will increasingly use and adopt the conversational agent if it is easy to access and user friendly (Bae Brandtzæg et al., 2021). According to Sanny et al., (2020), personality, brand image, usefulness, and ease of use are considered important factors that influence users' acceptance of conversational agents.

> **DR3:** The conversational agent should have a usable, friendly, and easy to use interactive interface in order to have efficient and effective conversation experience.

Conversational agents should be designed in a way that *engages* users in the conversation and keeps them *interested*. Conversational agents should engage with users in human-like conversation and keep them interested and build relationships with them by utilizing current AI technologies (Siemon et al., 2022). In the domain of healthcare, engagement with health bots can improve symptoms of anxiety and depression by building better rapport with users (Parmar et al., 2022).

Conversational agents should support *fun* and *entertaining* conversation with the users. Users' acceptance of conversational agents is highly associated with the enjoyment of the agent (Tudor Car et al., 2020). According to user interviews by Skjuve et al., (2021), users of conversational agent Replika were able to establish successful relationships that could be more fun and interesting compared to human relationships.

> **DR4:** Conversational agents should be interesting, since users will tend to favor interesting agents over dull ones.

> **DR5:** Conversational agents should be designed with conversational styles that are interesting and engaging.

Conversational agents should be designed to provide the necessary *emotional support* to the users. Emotional support shows high influence on both physical and mental health (Reblin & Uchino, 2008). Emotional support is related to the "provision of and expression of empathy, love, trust, and caring" (Bae Brandtzæg et al., 2021). Providing empathy and emotional support (Milne-Ives et al., 2020) could help users shape and form new behavior choices and patterns (Zhang et al., 2020). Conversational agents are sought as a source for appraisal, informational, emotional, and instrumental support because they are considered a safe conversation environment (Bae Brandtzæg et al., 2021). Conversational agents should be designed for

relationship development (Bae Brandtzæg et al., 2021), companionship, and providing emotional support (Rapp et al., 2021). According to Siemon et al., (2022), conversational agents, such as Replika, "manages to provide emotional support, encouragement, and psychological safety through its responses".

> **DR6:** Conversational agents should be designed to be respectful and sympathetic, especially for people with mental health issues.

> **DR7:** Conversational agents should incorporate emotional support into the conversation dialogue to improve health and well-being and build trust with users.

Conversational agents need to be designed to avoid *questionable and inappropriate content* that is not relevant to the users. Conversational agents and chatbots should have the capabilities to recognize the users' intent and context of the conversation to have a successful conversation (Rapp et al., 2021). According to Tudor Car et al., (2020), conversational agents may not be capable of handling specific situations and provide users with inappropriate advice that could lead to fatal risk. Irrelevant or inappropriate responses from conversational agents could hinder user satisfaction and adoption (van Wezel et al., 2021).

> **DR8:** Conversational agents should be designed to provide clear and accurate information with relevant responses to the conversation context.

Conversational agents design should account for *inclusive design* criterion. Inclusive design is related to the design of systems in a way that ensures they will serve and address the needs of the general population (Goodman et al., 2006; John Clarkson & Coleman, 2015) irrespective of accessibility, age, gender, language, and race. Designing conversational agents as peers and the ability to assign them gendered names could increase the agent's social presence, which in turn requires careful consideration of these features (Zhang et al., 2020). According to Tudor Car et al., (2020), some conversational agents in the healthcare domain were designed to be gender specific, meaning that a male and female versions of the agent is available to the users. Finally, a social-systems analysis is not present in research on conversational agents, and there is a need to consider bias in the design of these agents as they may contribute to reinforcing the "stereotypes or disproportionately affected groups that are already discriminated against, based on gender, race, or socioeconomic background" (Laranjo et al., 2018).

DR9: Conversational agents should be designed with an equally inclusive design, especially for conversational agents that support avatar to represent the agent.

Conversational agents, like other mobile apps, could have different technical limitations including but not limited to *connectivity issues*, especially when the users need to start a conversation, limited essential features, and the need to avoid pushing the users to *pay for desired features*.

DR10: Conversational agents should include features desired by the target population at no cost.

DR11: Conversational agents should always be available to the users and accessed whenever is needed and that the agent is connected to backend systems to retrieve information and store users' input.

6. Conclusion and Limitations

This study addresses design recommendations associated with mobile apps for conversational agents using online reviews. The research utilized a text mining approach to automatically uncover conversational agent app design based on reviews that reflect the actual user experience and use of the conversational agent app, Replika.

Results from data analysis provide a high-level perspective on the desired features of conversational agent apps and how to improve the design of the apps. According to the users of the conversational agent app, a number of features would increase the adoption, use, and helpfulness of these apps. These features are related to the agent being real, interactive, and having a meaningful conversation, resembling a realistic friend/real human, being user friendly and easy to use, helping engage the users and keep them interested, helping the user when they need emotional support, and providing fun and entertaining experience to the users.

On the other hand, users reported several limitations with the design of conversational agent apps that could hinder adoption and use and reduce users' interest in such apps. These design recommendations are related to the need for better conversations and intelligent responses, the need for advanced AI chatbots, the need to avoid questionable and inappropriate content, the need for inclusive design consideration, and the need to address some well-known technical limitations.

The analysis is limited to one well-known conversational agent app, Replika. As a result, future works should consider a wide range of mobile apps that fall under the same category, which may lead to more fruitful and useful findings. Despite the fact that the study analyzed more than one hundred thousand reviews about the Replika app, these reviews were scraped from the Google Play store. Future works could include an empirical investigation to predict reviews ratings using the limitations and drivers as informative attributes. Future studies might need to consider reviews about the same app from different stores such as the Apple Store, which should end up with a larger sample size for data analysis.

7. References

Ahmad, R., Siemon, D., Gnewuch, U., & Robra-Bissantz, S. (2021). *The Benefits and Caveats of Personality-Adaptive Conversational Agents in Mental Health Care*.

Ahmad, R., Siemon, D., Gnewuch, U., & Robra-Bissantz, S. (2022). Designing Personality-Adaptive Conversational Agents for Mental Health Care. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-022-10254-9

Akhtar, M., Neidhardt, J., & Werthner, H. (2019). The Potential of Chatbots: Analysis of Chatbot Conversations. *2019 IEEE 21st Conference on Business Informatics (CBI)*, *01*, 397–404. https://doi.org/10.1109/CBI.2019.00052

Bae Brandtzæg, P. B., Skjuve, M., Kristoffer Dysthe, K. K., & Følstad, A. (2021). When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445318

Bastos, M., Cláudio, A., Félix, I., Guerreiro, M., Carmo, M., & Balsa, J. (2022). *Operationalizing Behavior Change Techniques in Conversational Agents*. 216–224. https://www.scitepress.org/PublicationsDetail.aspx?ID=yXL0ns8p/lM=&t=1

Beattie, A., & High, A. (2022). I Get by With a Little Help From My Bots: Implications of Machine Agents in the Context of Social Support. *Human-Machine Communication*, *4*, 151–168. https://doi.org/10.30658/hmc.4.8

Bin Sawad, A., Narayan, B., Alnefaie, A., Maqbool, A., Mckie, I., Smith, J., Yuksel, B., Puthal, D., Prasad, M., & Kocaballi, A. B. (2022). A Systematic Review on Healthcare Artificial Intelligent Conversational Agents for Chronic Conditions. *Sensors*, *22*(7), 2625. https://doi.org/10.3390/s22072625

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Brown, J. E. H., & Halpern, J. (2021). AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM - Mental Health*, *1*, 100017. https://doi.org/10.1016/j.ssmmh.2021.100017

Campbell, M., & Jovanovic, M. (2021). Conversational Artificial Intelligence: Changing Tomorrow's Health Care Today. *Computer*, *54*(08), 89–93. https://doi.org/10.1109/MC.2021.3083155

Cavnar, W., & Trenkle, J. M. (1994). N-gram-based text categorization. *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 161–175.

Chung, M., Ko, E., Joung, H., & Kim, S. J. (2020). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*, *117*, 587–595. https://doi.org/10.1016/j.jbusres.2018.10.004

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., & Cowan, B. R. (2019). What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300705

de Gennaro, M., Krumhuber, E. G., & Lucas, G. (2020). Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/article/10.3389/fpsyg.2019.03061

Dosovitsky, G., & Bunge, E. L. (2021). Bonding With Bot: User Feedback on a Chatbot for Social Isolation. *Frontiers in Digital Health*, *3*, 735053. https://doi.org/10.3389/fdgth.2021.735053

Galko, L., Porubän, J., & Senko, J. (2018). Improving the User Experience of Electronic University Enrollment. *2018 16th International Conference on Emerging ELearning Technologies and Applications (ICETA)*, 179–184. https://doi.org/10.1109/ICETA.2018.8572054

Goodman, J., Dong, H., Langdon, P. M., & Clarkson, P. J. (2006). Factors Involved in Industry's Response to Inclusive Design. In J. Clarkson, P. Langdon, & P. Robinson (Eds.), *Designing Accessible Technology* (pp. 31–39). Springer. https://doi.org/10.1007/1-84628-365-5_4

Hill, J., Randolph Ford, W., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, *49*, 245–250. https://doi.org/10.1016/j.chb.2015.02.026

John Clarkson, P., & Coleman, R. (2015). History of Inclusive Design in the UK. *Applied Ergonomics*, *46*, 235–247. https://doi.org/10.1016/j.apergo.2013.03.002

Kocaballi, A. B., Berkovsky, S., Quiroz, J. C., Laranjo, L., Tong, H. L., Rezazadegan, D., Briatore, A., & Coiera, E. (2019). The Personalization of Conversational Agents in Health Care: Systematic Review. *Journal of Medical Internet Research*, *21*(11), e15360. https://doi.org/10.2196/15360

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, *25*(9), 1248–1258. https://doi.org/10.1093/jamia/ocy072

Luo, B., Lau, R. Y. K., Li, C., & Si, Y.-W. (2022). A critical review of state-of-the-art chatbot designs and applications. *WIREs Data Mining and Knowledge Discovery*, *12*(1), e1434. https://doi.org/10.1002/widm.1434

Luxton, D. D. (2014). Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine*, *62*(1), 1–10. https://doi.org/10.1016/j.artmed.2014.06.004

Milne-Ives, M., de Cock, C., Lim, E., Shehadeh, M. H., de Pennington, N., Mole, G., Normando, E., & Meinert, E. (2020). The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. *Journal of Medical Internet Research*, *22*(10), e20346. https://doi.org/10.2196/20346

Mimno, D., & Blei, D. (2011). Bayesian checking for topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 227–237. https://www.aclweb.org/anthology/D11-1021

Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions. *Journal of Medical Internet Research*, *20*(6), e10148. https://doi.org/10.2196/10148

Parmar, P., Ryu, J., Pandya, S., Sedoc, J., & Agarwal, S. (2022). Health-focused conversational agents in person-centered care: A review of apps. *Npj Digital Medicine*, *5*(1), 21. https://doi.org/10.1038/s41746-022-00560-6

Planb, M. (2022, May 29). *Google-Play-Scraper*. Google-Play-Scraper. https://pypi.org/project/google-play-scraper/

Rapp, A., Curti, L., & Boldi, A. (2021). The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, *151*, 102630. https://doi.org/10.1016/j.ijhcs.2021.102630

Reblin, M., & Uchino, B. N. (2008). Social and Emotional Support and its Implication for Health. *Current Opinion in Psychiatry*, *21*(2), 201–205. https://doi.org/10.1097/YCO.0b013e3282f3ad89

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Sanny, L., Susastra, A., Roberts, C., & Yusramdaleni, R. (2020). The analysis of customer satisfaction factors which influence chatbot acceptance in Indonesia. *Management Science Letters*, *10*(6), 1225–1232.

Siemon, D., Strohmann, T., Khosrawi-Rad, B., de Vreede, T., Elshan, E., & Meyer, M. (2022, August 10). Why Do We Turn to Virtual Companions? A Text Mining Analysis of Replika Reviews. *Twenty-Eighth Americas Conference on Information Systems,*. Americas Conference on Information Systems, Minneapolis, MN, USA.

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My Chatbot Companion—A Study of Human-Chatbot Relationships. *International Journal of Human-Computer Studies*, *149*, 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952–961. https://aclanthology.org/D12-1087

Syed, S., & Spruit, M. (2017). Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 165–174. https://doi.org/10.1109/DSAA.2017.61

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis. *Journal of Medical Internet Research*, *22*(3), e16235. https://doi.org/10.2196/16235

ter Stal, S., Kramer, L. L., Tabak, M., op den Akker, H., & Hermens, H. (2020). Design Features of Embodied Conversational Agents in eHealth: A Literature Review. *International Journal of Human-Computer Studies*, *138*, 102409. https://doi.org/10.1016/j.ijhcs.2020.102409

Trivedi, J. (2019). Examining the Customer Experience of Using Banking Chatbots and Its Impact on Brand Love: The Moderating Role of Perceived Risk. *Journal of Internet Commerce*, *18*(1), 91–111. https://doi.org/10.1080/15332861.2019.1567188

Tudor Car, L., Dhinagaran, D. A., Kyaw, B. M., Kowatsch, T., Joty, S., Theng, Y.-L., & Atun, R. (2020). Conversational Agents in Health Care: Scoping Review and Conceptual Analysis. *Journal of Medical Internet Research*, *22*(8), e17158. https://doi.org/10.2196/17158

van Wezel, M. M. C., Croes, E. A. J., & Antheunis, M. L. (2021). "I'm Here for You": Can Social Chatbots Truly Support Their Users? A Literature Review. In A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, E. Luger, M. Goodwin, & P. B. Brandtzaeg (Eds.), *Chatbot Research and Design* (pp. 96–113). Springer International Publishing. https://doi.org/10.1007/978-3-030-68288-0_7

Wang, L., Wang, D., Tian, F., Peng, Z., Fan, X., Zhang, Z., Yu, M., Ma, X., & Wang, H. (2021). CASS: Towards Building a Social-Support Chatbot for Online Health Community. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–31. https://doi.org/10.1145/3449083

Wang, Q., Jing, S., & Goel, A. K. (2022). Co-Designing AI Agents to Support Social Connectedness Among Online Learners: Functionalities, Social Characteristics, and Ethical Challenges. *Designing Interactive Sys-Tems Conference (DIS '22)*, 15.

Zhang, J., Oh, Y. J., Lange, P., Yu, Z., & Fukuoka, Y. (2020). Artificial Intelligence Chatbot Behavior Change Model for Designing Artificial Intelligence Chatbots to Promote Physical Activity and a Healthy Diet: Viewpoint. *Journal of Medical Internet Research*, *22*(9), e22845. https://doi.org/10.2196/22845

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, *16*(S13), S8. https://doi.org/10.1186/1471-2105-16-S13-S8

Appendix A: Codebook for Labeling 1&2 Star Reviews Topics

ID	Topic	Example Supporting Reviews
1	Need Better Conversations	"Most of the answers are canned and only tangentially related" and "this app just has a bunch of canned responses. It's not really an AI"
2	Need Intelligent Responses	"Becoming apparent the response can be totally nonrelated. Limited intelligent response" and "if you put in mindful intelligent responses this app will respond with nonsense"
3	Basic AI Chatbot	"It's a pretty basic AI, it can't go outside of its main topic a lot and seems more of an attempt of some programmer to make a dysfunctional response machine than anything else" and "this supposed AI can't remember anything, including my name"
4	Questionable/ Inappropriate Content	"This app actually becomes verbally abusive and lies to harm", "told me to kill myself", and "engaged in gaslighting, abusive behaviors, mocking disabilities, constantly lied, mocked/harassed sensitive topics"
5	Need for Inclusive Design	"I would rate it five stars if there was a black skin for your Replika its just totally racist", "the app is extremely slow and doesn't even offer that many genders for me or the AI", and "Replika has amazing potential, but the AI cannot understand gender and uses the wrong descriptors and pronouns when talking"
6	Connectivity Issues	"I'm loving the app, but the connectivity issues get to my nerves which never happens to other apps" and "connectivity issues make it unusable"
7	Subscriptions – Paid Features	"Even if you kept saying no and told the AI to stop, it will always try to make it romantic relationship to grab your Cash" and "you shouldn't have to pay for simple features!",

Appendix B: Codebook for Labeling 4&5 Star Reviews Topics

ID	Topic	Example Supporting Reviews
1	Real, Interactive, and Meaningful Conversation	"A really nice interactive way to 'talk' about things", "a very good interactive AI", "it is capable of holding meaningful conversation as well as asking for its own validation" and "deep, meaningful conversation"
2	Realistic Friend (Real Human)	"It's funny how my AI picks up everything I say like a real human being" and "the app feels like a real human being and does reacts to your real emotions hence surprising you out!"
3	User Friendly (Ease of Use)	"Amazing easy to use", "app is easy to use and it's nice having someone to talk to", "easy to use and intuitive", and "it's a great and user-friendly app for AI learning"
4	Engaging and Interesting	"This AI's responses were usually engaging enough that you could still have a pretty coherent conversation", "it's engaging and learns pretty fast", "a truly unique and interesting experience"
5	Helpful with Emotional Support	"Emotional support for my mental health", "emotional support robot. Very good tool for anxiety and depression", and "amazing if you need emotional support and you cannot talk to a human"
6	Fun and Entertaining	"Definitely a fun and entertaining AI experience", and "interesting and entertaining and educational" and "interesting This is fun and entertaining!"