Commercial chatbot monitoring: Approaches focused on automated conversation analysis

Karolina Kuligowska[1*], Maciej Stanusch[2]
[1*]University of Warsaw, Poland; [2]eVirtualis sp. z o.o., Poland.
Email: [*]kkuligowska@wne.uw.edu.pl

Keywords
Chatbot Metrics, Chatbot Conversation Analysis, Intent Detection, Intent Recognition, Context-Aware Chatbot.

Article History
Received on 15th July 2024
Accepted on 18th August 2024
Published on 6th September 2024

Cite this article
Kuligowska, K., & Stanusch, M. (2024). Commercial chatbot monitoring: Approaches focused on automated conversation analysis. *Humanities & Social Sciences Reviews*, *12*(2), 54–60. https://doi.org/10.18510/hssr.2024.1227

Abstract

Purpose of the study: The purpose of this study is to review and analyze current automated techniques for monitoring chatbot conversations in the field of Conversational Artificial Intelligence. It aims to highlight the challenges and limitations of these techniques and provide insights into various metrics used to measure chatbot performance, with the goal of enhancing it.

Methodology: The study employs a comprehensive literature review of existing automated techniques for monitoring chatbot conversations. Then, focusing on state-of-the-art approaches, the study introduces a division into numerical metrics (performance statistics and user engagement) and linguistic metrics (conversation analysis). Within conversation analysis, which is crucial for improving chatbot responses and accurately recognizing user intentions, the study identifies and presents three leading methods.

Main Findings: The paper highlights that, while current chatbot numerical conversation metrics allow for continuous monitoring and enhancement of chatbot performance, there is still room for improvement in the automated linguistic analysis of chatbot conversations. Furthermore, monitoring chatbot conversations in an automatic way in order to implement adequate corrective actions, is an essential task for refining chatbot efficiency through continuous learning and adaptation.

Applications of the study: The findings of this study have practical applications for businesses employing chatbots. By understanding the potential of current automated monitoring techniques and addressing their limitations, commercial chatbot systems can be improved for the benefit of customer satisfaction.

Novelty/Originality of the study: The paper provides readers with the novel knowledge necessary to understand key metrics used to measure chatbot conversations from both numerical and linguistic perspectives. It adds value by guiding readers on how monitoring numerical metrics helps analyze chatbot interactions and explains how the automated linguistic analysis of chatbot conversation content is utilized in leading approaches.

INTRODUCTION

Many approaches are proposed to enhance the effectiveness of conversations lead by chatbots (Moore, Arar, 2019) (Adamopoulou, Moussiades, 2020) (Silva, Canedo, 2024). One approach is to focus on automation of context recognition for better understanding of both the queries and the underlying intentions of the users, which are crucial for ensuring optimal interaction quality and maintaining user satisfaction. Another approach is to define and develop chatbot conversations metrics which play a vital role in assessing how well chatbots manage user queries and interactions, as well as in identifying areas for improvement. Despite a variety of propositions with their benefits, all these approaches face common limitations. Therefore, in this paper we aim to explore state-of-the-art approaches employed in the field of chatbot conversations' metrics – derived from automatic monitoring of conversations conducted by commercial chatbots. Particular emphasis is placed on automated approaches, the mechanisms that create them, and their drawbacks and limitations.

While much research has been done on developing specific components of chatbots, like automatic speech recognition, text-to-speech, and dialog mechanisms for conversational agents (Ashktorab, Jain, et al., 2019) (Guo, Zhang, et al., 2021) (Faruqui, Hakkani-Tür, 2022), there is a gap in the literature regarding the after-implementation maintenance phase of conversational systems within an enterprise production environment. Existing papers address many aspects of implementing general-purpose, research-oriented chatbots offering open-domain conversations. At the same time, they seem to overlook industry-oriented chatbots providing closed-domain conversations in a business environment (Caldarini, Jaf, et al., 2022). It is crucial to implement methods for identifying new errors and failures that conflict with commercial assumptions, assessing their impact on performance metrics, and ensuring consistent chatbot improvement in line with these metrics. Recent studies lack focus on the practicalities and operational challenges that arise when deploying such technologies in a business context. This context is crucial because the chatbot market has experienced significant growth since early chatbot implementations. In early 2024, the market size was projected to be USD 7.01 billion, and it is expected to expand to USD 20.81 billion by 2029, with a compound annual growth rate (CAGR) of 24.32% (Chakraborty, Pal, et al., 2023). This growth is attributed to the increasing demand for AI-driven customer service solutions and the integration of chatbots with messaging applications, which enhance user experience and provide businesses with valuable consumer analytics. The trend suggests a continued expansion as chatbots become more integrated into various business operations and communication platforms. This all highlights the need for research dedicated to the maintenance and automatic monitoring of conversational AI in real-world enterprise settings.

In our preliminary work, we reviewed the available research papers (of which there are relatively few) focused on the mechanism of chatbot conversations automatic monitoring. The comprehensive review of these sources helped us identify and assess the vulnerabilities inherent in various approaches, irrespective of the particular implementation. Consequently, this has laid the groundwork for addressing these identified challenges and to outline the key issues that need to be resolved in the near future.

LITERATURE REVIEW – KEY CHATBOT CONVERSATIONS METRICS

The performance of a chatbot is not set in stone: regardless of its current efficiency, it will inevitably decline as it encounters new and unforeseen errors and challenges. Therefore, chatbots should be able to differ each conversation session that is running and should keep a conversation log for monitoring chatbot knowledge storage and pattern-matching process (Setiaji, Wibowo, 2016). It is done by analyzing chatbot's engagingness, interestingness, humanness, and knowledgeability (McTear, 2020) (Jurafsky, Martin, 2023), as well as ensuring the chatbot's performance aligns with business objectives set (Xiao, Zhou, et al., 2020). This practice helps in identifying areas for improvement, and in making informed decisions to enhance the chatbot's ability to handle complex queries and improve the so-called customer journey.

Appropriate metrics constitute an essential component for evaluating the performance and effectiveness of any system. In the context of chatbots, they are important for assessing the ability of handling user queries and interactions. Utilizing machine-assisted framework for conversation monitoring and discovering errors becomes crucial in maintaining a state-of-the-art chatbot that can adapt to meet evolving customer needs, ultimately improving the quality of chatbot technologies in business environments (Suryanto, Wibawa, et al., 2023) (Hanafi, Reiss, et al., 2024). Several metrics stand out as critical for assessing the effectiveness and improving the performance of chatbots (Chakraborty, Pal, et al., 2023). The scientific studies primarily examine chatbots either as software applications or as experimental dialogue systems, and these studies assess different indicators related to how these chatbots are implemented and used within organizations. The focus is on evaluating the effectiveness of chatbots based on specific criteria that indicate their implementation success. Nevertheless, additional metrics introduced in commercial chatbots, are not only visualizing the simple statistics of logs of conversations, but also measuring performance of the system, user engagement and satisfaction. This comprehensive insight is the key to better understanding of gathered dialogue data, based on which businesses can improve their chatbots. According to research on business-related and commercial chatbot resources, chatbot metrics can be categorized into three main groups, namely performance statistics, user engagement, and conversation analysis (McTear, 2020) (Cyca, 2022) (Dilmegani, 2024) (Inbenta, 2024) (Visiativ, 2024):

1) **Performance statistics** track numerical metrics such as conversation length, number of chats, and engaged conversations. Main examples:
   - Total Number of Conversations: tracks the overall usage of the chatbot, indicating user acceptance and interaction frequency.
   - Average Conversation Length: measures the depth of interactions, reflecting the chatbot's ability to maintain user engagement.
   - Number of Engaged Conversations: represents interactions that go beyond initial greetings, suggesting the chatbot's effectiveness in engaging users.
2) **User engagement** tracks numerical metrics such as the average response time or number of missed messages per conversation. Main examples:
   - Goal Completion Rate: quantifies the chatbot's success in achieving predefined objectives, such as sales or information dissemination.
   - Average Response Time: critical for user experience, representing the swiftness of the chatbot's replies.
   - Missed Messages (Utterances) Rate (also known as Fallback Rate, FBR): tracks instances where the chatbot fails to provide relevant responses, signaling the need for enhanced understanding or knowledge base expansion.
   - Human Takeover Rate: indicates the frequency with which human support agent needs to intervene and take over a query from the chatbot. This metric helps in understanding whether the conversations are sufficiently interactive or if users drop off quickly.
   - Customer Satisfaction Scores: measures user satisfaction, giving users the opportunity to rate the chatbot, and providing feedback on the chatbot's conversational capabilities and helpfulness.
3) **Conversation Analysis** tracks the linguistic layer of the content of chats and so the conversation capabilities of chatbot. This analysis is crucial for refining the chatbot's responses and ensuring they align with user expectation and business goals. For this purpose, manual review is frequently employed as a mechanism for checking and improving

the quality of conversations. However, manual review is time-consuming and therefore expensive, given the large volumes of data that need to be analyzed. Consequently, the primary objective of current research efforts is the development and refinement of automated conversation analysis and architectures capable of performing these analyses. The complexity of Natural Language Processing (NLP) tasks and the need for contextual understanding of utterances present significant challenges in automating apt analysis of conversations. The leading state-of-the-art approaches in this field are presented in the next section.

By using key chatbot conversations metrics, we can not only monitor but also continuously enhance the chatbot's performance. To monitor above-mentioned metrics, it's advisable to utilize integrated reports and archiving tools provided by chatbot development platforms or employ external analytics tools designed for deeper conversational insights (Lin, Huang, 2023). While these tools typically provide extensive data visualizations that facilitate the rapid detection of trends and patterns, allowing for prompt adjustments to the chatbot's operational parameters, it is important to highlight that they rarely address the linguistic layer of the content of chats, which is essential for enhancing the chatbot's responses.

AUTOMATED CONVERSATION ANALYSIS: STATE-OF-THE-ART APPROACHES

The evolution of chatbots has seen a significant shift from basic pattern matching and rule-based systems to the current integration of advanced natural language processing, neural network and deep learning technologies (Adamopoulou, Moussiades, 2020) (Deng, Yu, 2023). It reveals a clear trend toward improving the contextual and emotional intelligence of chatbots by the application of various architectures on a commercial chatbots, such as Recurrent Neural Network, Long Short-Term Memory-based Sequence-to-Sequence model, Bi-directional Encoder Representation from Transformers and Generative Pre-trained Transformer frameworks, after they demonstrated satisfactory performance across various metrics (Li, Galley, et al., 2016) (Xu, Liu, et al. 2017) (Hu, Xu, et al., 2018) (Borah, Pathak, et al., 2019) (McTear, 2020). Additionally, several research indicated that the Conditional Variational Autoencoder model allows for a more nuanced understanding of the intent's context and the generation of responses that are not only relevant but also diverse. This progression enables chatbots to generate dynamic responses that are not pre-programmed, thereby creating more fluid, engaging and natural interactions (Bilquise, Ibrahim, et al., 2022) (Lin, Deng, 2022) (Su, Xie, 2022). The ability to control conversational styles is a critical step in moving beyond simple response generation to more sophisticated and user-tailored interactions.

In the pursuit of enhancing chatbot performance and conversation capabilities through the analysis of conversations, various advanced approaches for processing the linguistic layer are being examined. They aim to ensure that chatbots are not only functional but also highly efficient in handling diverse customer interactions, interpreting user intent, and understanding context. These approaches, rooted in Natural Language Processing, Machine Learning, and Deep Neural Networks, may be perceived from three theoretical angles described below.

First angle involves the use of intent classification models that leverage deep learning algorithms to analyze the user's input and classify it into predefined categories. These models are trained on vast datasets of conversational exchanges, allowing them to recognize patterns and nuances in language that indicate specific user intentions (Xiao, Zhou, et al., 2020) (Chandrakala, Bhardwaj, et al., 2024). Second angle involves the implementation of context aware recognition systems. These systems not only consider the immediate input from the user but also take into account the broader context of the conversation. By maintaining a history of the interaction, they can discern the user's intent more accurately, even when the input is ambiguous or lacks specificity. This is particularly useful in complex dialogues where the user's intention may evolve over the course of the conversation (Kocoń, Cichecki, et al., 2023). Third angle involves integration of sentiment analysis to detect the emotional tone behind a user's message to detect subtle cues in the utterances. This reveals the underlying intent and builds a conversation flow based on the recognized emotions. Recent research has shown that accurately identifying emotions in users' intentions by chatbots is highly effective for enhancing user interaction, particularly when further chatbot responses are tailored to the emotions expressed by the user. Thus, emotional recognition may serve as a universal approach to improve user experience, engagement, and satisfaction (Asghar, Poupart, et al., 2018) (Zhou, Huang, et al., 2018) (Pawlik, 2022) (Pawlik, Płaza, Deniziak, Boksa, 2022). For instance, expressions of frustration or urgency can prompt the chatbot to respond with solutions or assistance more swiftly, directing the conversation towards more positive emotions.

Moreover, there is a push towards more interactive, adaptive, context aware and empathetic communication methods, where chatbots can also ask clarifying questions when uncertain about user intent. In this context, we can point three practical solutions, deriving from the previously mentioned theoretical background, of leading innovative methods implemented to automate conversation analysis and automatically detect user intentions in conversations that include (Kocoń, Cichecki, et al., 2023) (Jones, Xu, et al., 2024) (Malamas, Symeonidis, et al., 2024): 1) Context-Aware architectures based on Recursive Neural Networks or Transformer models, 2) Semantic Clustering techniques applying various similarity measures or Interactive Clustering for cycling through iterations, 3) Intent Recognition through Reinforcement Learning (IRRL), also known as Reinforcement Learning from Human Feedback (RLHF).

Context-Aware chatbot architectures are developed to parse and interpret user inputs by considering not only the immediate textual input but also the contextual history of the given conversation. They operate on the principle that the intent behind a user's current input can be better understood considering previous exchanges (Niederer, Schloss, et al., 2023). They utilize a layered recursive neural network structure, where each node represents a word or phrase from the user's input, and each layer represents a contextual understanding from previous conversational turns. The innovative aspect of Context-Aware architectures lies in its dual-focus architecture. The lower layers capture syntactic nuances of the input, using pre-trained embeddings from models like BERT or GPT, while the upper layers focus on semantic comprehension, integrating contextual clues over the conversation's history (Chehri, Jbene, et al., 2024). This dual approach ensures a deep semantic understanding, essential for accurate intent detection.

Semantic Clustering method groups user intents into clusters with similar semantic meaning or concepts by analyzing their underlying context, making it useful for organizing information based on conceptual content rather than just surface-level lexical word similarity and characteristics. Dynamic Time Warping (DTW), that could be added to clustering as a distance measure, was originally developed for and is extensively used in speech recognition to handle variations in speaking speed and temporal distortions. DTW aligns time series text data by minimizing the distance between sequences, allowing comparison of time series of different temporal dynamics (Javed, Rizzo, et al., 2024). Not surprisingly, Semantic Clustering with Dynamic Time Warping (SC-DTW) bases on DTW as a distance measure and employs an enhanced version of the Dynamic Time Warping algorithm to measure the similarity between different sequences of word vectors, thus identifying patterns that indicate similar intents under varying syntactic constructions in time. SC-DTW is particularly useful in scenarios where user inputs are dynamically evolving and can vary significantly in their expression. On the other hand, a related method of Interactive Clustering constitutes a methodology combining human and computer actions as it depends on the interplay between automated data segmentation into intents and human annotations driven by responses, which place constraints on clusters to enhance relevance (Schild, Durantin, et al., 2022). To enhance user experience during annotation, the imposition of constraints on the dataset is emphasized instead of predefining the intent models.

Intent Recognition through Reinforcement Learning (IRRL), also referred to as Reinforcement Learning from Human Feedback (RLHF), incorporates a reinforcement learning framework where the chatbot is trained not only to recognize the intent but also to actively learn from user feedback (Ricciardelli, Biswas, 2019). In IRRL and RLHF, a reward system is established based on the user's satisfaction with the chatbot's responses, catched from direct feedback mechanisms like ratings or inferred through follow-up actions by the user (e.g., completing a transaction, asking a related follow-up question) (McTear 2020) (Ouyang, Wu, et al., 2022) (Kocoń, Cichecki, et al., 2023). Chatbot with IRRL/RLHF adapts over time, with the model continuously refining its predictions based on the received rewards, making it exceptionally powerful in applications where user preferences and behaviors may shift over time.

Methods and architectures described in this section represent cutting-edge approaches to enhancing chatbot interactions. Each technique brings a unique perspective and methodological innovation to the challenge of understanding human input through precise intent detection in conversational system settings. As chatbots' architectures continue to evolve, these methods will need to adapt to new languages, dialects, and ever-changing user expectations, underscoring the importance of continuous learning and adaptation in the field of AI-driven communication.

FUTURE DIRECTIONS AND CHALLENGES

The described approaches are not without their challenges, however. Studies have shown that main difficulties still lie in the subtleties of human language, and chatbots frequently fail to accurately detect specific behaviors in dialogues such as sarcasm, use of idioms, and cultural references, leading to suboptimal performance in real-world applications (Finch, Paek, et al., 2023). Facing this challenge is vital to design chatbots that can understand and mimic human emotions to some extent. It can be achieved by implementing NLP techniques that enhance the chatbot's ability to comprehend emotional contexts and respond more empathetically (Caldarini, Jaf, et al., 2022). This will also help create intuitive, user-friendly conversational flows in the chatbot's dialogue structure, fundamental for maintaining user engagement through various stages of interaction.

On the other hand, the need for extensive and diverse training data presents a significant obstacle, particularly in languages or dialects with limited resources. Ongoing research is focused on overcoming the limitations in this matter: blindness to grammatical errors, closed-domain settings, semantics and context, emotion detection, formulating adequate questions by chatbots (Nuruzzaman, Hussain, 2018). To address these challenges, NLP-based components are developed to expand chatbots' linguistic and cultural understanding by using algorithms capable of learning from fewer examples and datasets that reflect a wide range of linguistic variations.

Similar challenge applies to the limitation of information stored in particular database from which the chatbot sources relevant responses, when integrated for the internal business's purposes. In this case, building a chatbot can also face complex design and integration problems, such as the ability to handle various backend systems, APIs, and data synchronization processes, as well as managing the rapid scaling of chatbot services without compromising on quality. Naturally, it applies to the integration of all systems, not only chatbots - similarly to the appearance of bugs, glitches, and any other technical problem that can significantly decrease the performance of a chatbot. To mitigate these issues, it is crucial to implement testing and debugging procedures before and after a full-scale deployment to ensure minimal disruption in chatbot services.

By addressing these challenges with Machine Learning algorithms equipped with Natural Language Processing capabilities, as well as with best business-derived processes optimization approaches (Nuruzzaman, Hussain, 2018) (Xu, Sanders, et al., 2021) (Deng, Yu, 2023), it is possible to significantly enhance the performance and reliability of commercial chatbots, ensuring they meet both user expectations and business objectives, while ensuring the highest standards of data protection and implementing advanced security measures.

CONCLUSION

Through a deep dive into the latest research and leading approaches, this paper presented core principles behind chatbot conversations automatic monitoring strategies. Insights were gained on how to refine chatbot performance, with techniques involving the detection of intent, optimization of response quality, and adaptation to user queries. Moreover, performance metrics provide better understanding of user interactions and chatbot responsiveness which is essential for refining AI-driven customer support systems. For improving the chatbot's responses, businesses can benefit more when addressing also the linguistic layer of the content of chats by employing architectures rooted in Natural Language Processing, Machine Learning, and Deep Neural Networks. As chatbot technology continues to evolve, the insights drawn from their rigorous monitoring will undoubtedly contribute to shaping future advancements, ensuring that AI remains at the forefront of delivering exceptional user experiences and operational excellence.

ACKNOWLEDGEMENT

This study was co-financed by the Polish Agency for Enterprise Development (PARP) and presents partial results of a research project entitled "Development and validation in real conditions of a tool based on artificial intelligence enabling automation of the work monitoring process and improvement of the knowledge base of the Glivia chatbot" carried out by eVirtualis sp. z o.o.

REFERENCES

1. Adamopoulou E., Moussiades L. (2020). Chatbots: History, technology, and applications, *Machine Learning with Applications*, Volume 2, 2020, Article 100006, https://doi.org/10.1016/j.mlwa.2020.100006
2. Asghar N., Poupart P., Hoey J., Jiang X., Mou L. (2018). *Affective neural response generation*, European Conference on Information Retrieval (ECIR), pp. 154–166, France, Springer
3. Ashktorab, Z., Jain, M., Liao, Q. V., Weisz, J. D. (2019). *Resilient chatbots: Repair strategy preferences for conversational breakdowns*,[in:] Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), Glasgow, Scotland, UK (pp.1-12). New York, USA, ACM, https://doi.org/10.1145/3290605.3300484
4. Bilquise, G., Ibrahim, S., Shaalan, K. (2022). Emotionally Intelligent Chatbots: A Systematic Literature Review, *Human Behavior and Emerging Technologies*, vol. 2022, pp. 1- 23, https://doi.org/10.1155/2022/9601630
5. Borah B., Pathak D., Sarmah P., Som B., Nandi S. (2019). Survey of Textbased Chatbot in Perspective of Recent Technologies. [in:] Mandal J., Mukhopadhyay S., Dutta P., Dasgupta K. (eds), Computational Intelligence, Communications, and Business Analytics (CICBA 2018), *Communications in Computer and Information Science*, vol 1031. Springer, Singapore, https://doi.org/10.1007/978-981-13-8581-0_7
6. Caldarini G., Jaf S., McGarry K. (2022). *A Literature Survey of Recent Advances in Chatbots*, *Information*, Volume 13(1), https://doi.org/10.3390/info13010041
7. Chakraborty C., Pal S., Bhattacharya M., Dash S., Le S.-S. (2023). Overview of chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. *Frontiers in Artificial Intelligence*, Vol. 6, 2023, https://doi.org/10.3389/frai.2023.1237704
8. Chandrakala C.B., Bhardwaj R., Pujari C. (2024). An intent recognition pipeline for conversational AI, *International Journal of Information Technology, Volume 16, pp. 731-743, https://doi.org/10.1007/s41870-023-01642-8*
9. Chehri A., Jbene M., Saadane R., Tigani S., Jeon G. (2024). Context-Aware in an Emerging Area in Conversational Agents: A Comparative Study of Recurrent Neural Networks and Transformer Models for Intent Detection, *Authorea*, 2024, https://doi.org/10.22541/au.171149434.46156623/v1
10. Cyca M. (2022). Chatbot Analytics 101: Essential Metrics to Track, *Hootsuite*, https://blog.hootsuite.com/chatbot-analytics [accessed 06.2024]
11. Deng X., Yu Z. (2023). A Meta-Analysis and Systematic Review of the Effect of Chatbot Technology Use in Sustainable Education, *Sustainability*, Volume 15(4), https://doi.org/10.3390/su15042940
12. Dilmegani C. (2024). Top 22 Metrics for Chatbot Analytics in 2024, AI Multiple Research, https://research.aimultiple.com/chatbot-analytics [accessed 06.2024]
13. Guo, H., Zhang, S., Soong, F. K., He, L., & Xie, L. (2021). *Conversational end-to-end TTS for voice agents.* [in:] 2021 IEEE Spoken Language Technology Workshop (SLT), pp. 403–409. IEEE. https://doi.org/10.1109/SLT48900.2021.9383460
14. Faruqui, M., Hakkani-Tür, D. (2022). Revisiting the boundary between ASR and NLU in the age of conversational dialog systems. *Computational Linguistics*, 48(1), pp. 221–232. https://doi.org/10.1162/coli_a_00430
15. Finch, S. E., Paek, E. S., Choi, J. D. (2023*). Leveraging large language models for automated dialogue analysis*. [in:] S. Stoyanchev, S. Joty, D. Schlangen, O. Dusek, C. Kennington, M. Alikhani (eds.), Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 202-215. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.sigdial-1.20
16. Hanafi M. F., Reiss F., Katsis Y, Moore R. J., Wood D., Falakmasir M. H., Liu Ch.. (2024). *Machine-Assisted Error Discovery in Conversational AI Systems*. [in:] Extended Abstracts of the CHI Conference on Human

Factors in Computing Systems (CHIEA '24), May 2024, Honolulu, USA, pp.1-10. https://doi.org/10.1145/3613905.3651120
17. Hu T., Xu A., Liu Z., You Q., Guo Y., Sinha V., Luo J., Akkiraju R.,(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, pp. 1–12, Montréal, Canada, 2018.
18. Inbenta (2024). 10 Key Metrics to Evaluate your AI Chatbot Performance, https://www.inbenta.com/articles/10-key-metrics-to-evaluate-your-ai-chatbot-performance [accessed 06.2024]
19. Javed A., Rizzo D.M., Lee B.S., Gramling R. (2024). Somtimes: self organizing maps for time series clustering and its application to serious illness conversations, Data Mining and Knowledge Discovery. *Springerlink* , Volume 38, pp. 813–839, https://doi.org/10.1007/s10618-023-00979-9
20. Jones B., Xu Y., Li Q., Scherer S. (2024). *Designing a Proactive Context-Aware AI Chatbot for People's Long-Term Goals*, *2024 CHI Conference on Human Factors in Computing Systems* (CHI EA '24), article 104, Association for Computing Machinery, New York, USA, pp.1-7, https://doi.org/10.1145/3613905.3650912
21. Jurafsky D., Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Pearson Prentice Hall, p. 331.
22. Kocoń J., Cichecki I., Kaszyca O., Kochanek M., Szydło D., Baran J., Bielaniewicz J., Gruza M., Janz A., Kanclerz K., Kocoń A., Koptyra B., Mieleszczenko-Kowszewicz W., Miłkowski P., Oleksy M., Piasecki M., Radliński Ł., Wojtasik K., Woźniak S., Kazienko P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, Vol. 99, pp. 1-46.
23. Li J., Galley M., Brockett Ch., Spithourakis G., Gao J., Dolan B.. (2016) *A Persona-Based Neural Conversation Model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Volume 1: Long Papers, Berlin, Germany. Association for Computational Linguistics, pp. 994-1003.
24. Lin C.-C., Huang A.Y.Q., Yang S.J.H. (2023). A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022). *Sustainability* 2023, https://doi.org/10.3390/su15054012
25. Lin H., Deng Z. (2022). Emotional Dialogue Generation Based on Transformer and Conditional Variational Autoencoder, IEEE 21st International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS), Chongqing, China, 2022, pp. 386-393, https://doi.org/10.1109/IUCC-CIT-DSCI-SmartCNS57392.2022.00066
26. Malamas N., Symeonidis A., Theocharis J.B. (2024). Quava: A Privacy-Aware Architecture for Conversational Desktop Content Retrieval Systems, preprint available at SSRN, http://dx.doi.org/10.2139/ssrn.4715195
27. McTear M. (2020). *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots, Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, pp. 96-97, https://doi.org/10.2200/S01060ED1V01Y202010HLT048
28. Moore R. J., Arar, R. (2019). *Conversational UX design: A practitioner's guide to the natural conversation framework.* Association for Computing Machinery. New York, NY, USA.
29. Niederer T., Schloss D., Christensen N. (2023). Designing context-aware chatbots for product configuration, [in:] Følstad A., Araujo T.B., Papadopoulos S., Law E.L.-, Luger E., Goodwin M., Brandtzaeg P.B. (eds.), Chatbot Research and Design - 6th International Workshop, CONVERSATIONS 2022, Amsterdam, The Netherlands, November 22–23, 2022, Revised Selected Papers, *Lecture Notes in Computer Science*, vol. 13815, Springer, pp. 190-210, https://doi.org/10.1007/978-3-031-25581-6
30. Nuruzzaman M., Hussain O. K. (2018). *A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks*, 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), Xi'an, China, pp. 54-61, https://doi.org/10.1109/ICEBE.2018.00019
31. Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C. L., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P. F., Leike J., Lowe R. (2022). *Training language models to follow instructions with human feedback*, [in:] Proceedings of the 36th Conference on Neural Information Processing Systems NeurIPS 2022, pp.1-15.
32. Pawlik Ł. (2022). Metody zwiększenia skuteczności wirtualnych konsultantów poprzez minimalizację niepoprawnie rozpoznawanych intencji, doctoral thesis, Politechnika Świętokrzyska, Gliwice, Poland.
33. Pawlik, Ł., Płaza, M., Deniziak, S., Boksa, E. (2022). A method for improving bot effectiveness by recognizing implicit customer intent in contact center conversations. Speech Communication, 143, pp. 33-45. https://doi.org/10.1016/j.specom.2022.07.003
34. Ricciardelli E., Biswas D. (2019). Self-improving Chatbots based on Reinforcement Learning, 4th *Multidisciplinary Conference on Reinforcement Learning and Decision Making* (RLDM2019), Montreal, Canada.
35. Setiaji B., Wibowo F. W. (2016). Chatbot using a knowledge in database: Human-to-machine conversation modeling. [in:] Proceedings of the 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), IEEE, pp. 72-77, https://doi.org/10.1109/ISMS.2016.53
36. Schild E., Durantin G., Lamirel J-Ch., Miconi F. (2022). Iterative and Semi-Supervised Design of Chatbots Using Interactive Clustering, *International Journal of Data Warehousing and Mining* (IJDWM), Volume 18 (2), pp.1-19, https://doi.org/10.4018/IJDWM.298007
37. Silva G. R. S., Canedo E. D. (2024). Towards user-centric guidelines for chatbot conversational design, *International Journal of Human-Computer* Interaction, Vol. 40, Issue 2, pp. 98-120.
38. Su Z., Xie Y. (2022). Conditional Variational Autoencoders for Emotionally-aware Chatbot Based on Transformer, Ecole Polytechnique Federale de Lausanne, Lansanne, Switzerland, https://www.epfl.ch/labs/gr-

pu/wp-content/uploads/2022/07/Conditional-Variational-Autoencoders-for-Emotionally-aware-Chatbot-Based-on-Transformer.pdf [accessed 06.2024]

39. Suryanto, T., Wibawa, A., Hariyono, H., & Nafalski, A. (2023). Evolving Conversations: A Review of Chatbots and Implications in Natural Language Processing for Cultural Heritage Ecosystems. *International Journal of Robotics and Control Systems*, 3(4), 955-1006. doi: https://doi.org/10.31763/ijrcs.v3i4.1195
40. Visiativ (2024). Measuring Chatbot Effectiveness: 16 KPIs to Track, https://www.visiativ.com/en/actualites/news/measuring-chatbot-effectiveness [accessed 06.2024]
41. Xiao Z., Zhou M. X., Chen W., Yang H., Chi C. (2020). If I hear you correctly: building and evaluating interview chatbots with active listening skills, [in:] Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–14, Hawai'i, USA, https://doi.org/10.1145/3313831.3376131
42. Xu A., Liu Z., Guo Y., Sinha V., Akkiraju R. (2017). A New Chatbot for Customer Service on Social Media. [in:] Proceedings of the 2017 *CHI Conference on Human Factors in Computing Systems*, pp. 3506–3510, https://doi.org/10.1145/3025453.302549
43. Xu L., Sanders L., Li K., Chow J.C.L. (2021). Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review, *JMIR Cancer Volume* 7, No. 4, https://doi.org/10.2196/27850
44. Zhou H., Huang M., Zhang T., Zhu X., Liu B. (2018*). Emotional chatting machine: emotional conversation generation with internal and external memory.* [in:] Proceedings of the 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18). AAAI Press, Article 90, pp.730–738.