IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

Chatbot Performance Evaluation

Kanchan Katake[1], Sanchay Kumar[2], Divya Nair[3], Rutuja Kadam[4], Aishwarya Kshirsagar[5]
Assistant Professor, Department of IT[1]
Students, Department of IT[2,3,4,5]
MIT Art, Design & Technology, Pune, India

Abstract: *Intelligent conversational computer programs, commonly known as chatbots, have become a popular means of providing automated online assistance and guidance. Despite being computer programs, chatbots are designed to simulate human-like conversation and provide users with the impression that they are interacting with a human. This has made chatbots a widely adopted tool for virtual customer support across many businesses. Chatbots employ Artificial Intelligence, particularly Natural Language Processing and Machine Learning, to operate effectively. However, the use of chatbots is not without its challenges and limitations. The survey begins by reviewing the background of chatbots, including their history, definition, and applications. This is followed by an overview of Artificial Intelligence and its role in chatbot development. The paper then examines the challenges and limitations associated with chatbots, including their limited ability to understand complex user inquiries and the potential for chatbots to make errors. The survey also provides an analysis of current research trends in chatbot development, including approaches to Natural Language Processing, Machine Learning, and user interface design. The paper identifies key gaps in knowledge related to chatbot design, particularly in the areas of user interaction, context awareness, and response generation. The survey concludes by offering recommendations for future research in chatbot development. These include the need for research on chatbots' ability to understand complex user inquiries, the development of new algorithms for context awareness and response generation, and the exploration of new user interface designs for chatbots. Through this survey, we hope to provide valuable insights for researchers and practitioners in the field of chatbot development, and to stimulate new avenues of research in this exciting and rapidly evolving field.*

Keywords: Chatbots, Artificial Intelligence, Natural Language Processing, User Satisfaction, Performance Evaluation, Machine Learning

I. INTRODUCTION

Chatbots have emerged as valuable tools across industries, providing efficient and automated customer interactions. However, understanding the factors that influence chatbot performance and user experience is critical for optimizing their effectiveness. By evaluating existing chatbots, this study aims to identify and address the defects that impact user satisfaction and explore strategies to restructure and enhance chatbot performance. Insights gained from this research will help businesses in diverse sectors leverage chatbot analytics and metrics to gauge performance accurately.The motivation behind this project lies in the need to evaluate the chatbots available online and comprehensively understand how defects affect user experiences. Surveys conducted with businesses utilizing chatbots have shown varying metrics employed across different industries to measure performance. For instance, in the banking and financial sectors, chatbots primarily focus on improving user efficiency, reducing call volumes, and minimizing service costs. By studying chatbot evaluation methods, we aim to incorporate the findings into the development of our own chatbot, specifically designed for a domain-specific environment.Despite significant progress in chatbot development, there remains a lack of research on the most effective metrics for evaluating chatbot performance and the influence of user experience on their effectiveness. This study aims to bridge this gap by conducting a comparative analysis of performance metrics and user experience. Through this investigation, we seek to identify the key evaluation criteria that reliably assess chatbot performance and understand how user experience can be improved to enhance chatbot

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

effectiveness. The outcomes of this research will provide valuable insights for the development and deployment of chatbots, contributing to the advancement of the conversational AI field.The primary objective of chatbot performance evaluation is to assess the quality and effectiveness of chatbots in conversing with human users. This includes evaluating their ability to understand and respond to user inputs, as well as providing accurate and useful information. By identifying areas for improvement and optimizing chatbot performance, the goal is to enhance the overall user experience. Additionally, the knowledge gained from working with diverse chatbot types available online will inform the creation of our own chatbot, tailored to a specific domain.

II. LITERATURE REVIEW

A. In the paper "Commercial Chatbot: Performance Evaluation, Usability metrics and quality assurance" [1] Karolina Kuligo wska states that the relevance of performance, usability, and overall quality evaluation of every commercial application of virtual assistant was illustrated by the 10 factors that determine the quality of every commercial chatbot deployment. The ten factors taken into account are: a speech synthesis unit, built-in knowledge base, conversational skills, and content sensitivity.

B. In the research paper "Algorithm Inspection for Chatbot Performance Evaluation" [2] Vijayaraghavan V, Jack Brian Cooper and Rian Leevinson J study the strategies that can be utilized to evaluate chatbot performance are the focus of this research. Testing chatbot output: Chattest (120 questions: Answering, Error Management, Intelligence, Navigation, Onboarding, Personality and Understanding). In Algorithm examination we can consider Naive Bayes, SVM and Natural Language Processing (NLP): this includes Grammar and Parsing Algorithms, Statistical Parsing, Verification and Validation.

C. In the paper "Enhancing community interaction with data driven chatbot- DBpedia Chatbot" [3] Ram G Athreya, Ricardo Usbeck and Axel-Cyrille Ngomo study about knowledge graph driven chatbot which is created to maximize community interaction, where the authors describe the internal working which is TF-IDF vectorization, K-means Clustering, rule-based dialogue, query fulfillment and Natural Language Processing questions.

D. The research paper "A Literature Survey of Recent Advances in Chatbots" [5] Guendalina Caldarini, Sardar Jaf and Kenneth McGarry provides an overview of recent advancements in chatbot technology. The author explores the evolution of chatbots, their current state, and the challenges they face. The paper also discusses the role of natural language processing (NLP) in chatbots, and how it has improved their performance. The author highlights the importance of user experience and provides insights into future directions for chatbot research, including improving their intelligence and increasing their capabilities.

E. The research paper "The Impact of Chatbots in Healthcare Processes: A Systematic Review," [7] by J. Haun et al , reviews the existing research on the use of chatbots, based on objective measures such as response time, message length, and user satisfaction ratings. The authors also suggest using machine learning algorithms to analyze chat logs and identify areas for improvement. The proposed methodology was tested through an experiment with a chatbot designed to provide information about local restaurants, showing promising results in identifying areas for improvement and measuring chatbot performance more objectively..

III. METHODOLOGY

Research Design

The initial step in evaluating chatbot performance is to determine the research design, which offers two options: a userbased evaluation or a task-based evaluation. A user-based evaluation aims to measure user satisfaction regarding the chatbot's overall performance and user experience. It involves assessing how well the chatbot meets user needs and expectations. Feedback from users through surveys, interviews, or rating scales is collected to gauge satisfaction levels, ease of use, and perceived helpfulness. On the other hand, a task-based evaluation focuses on assessing the chatbot's performance in successfully completing specific tasks. It emphasizes the chatbot's functional capabilities rather than user experience. Performance metrics such as task completion rate, response accuracy, and response time are used to

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

evaluate the chatbot's proficiency. Choosing between user-based and task-based evaluations depends on research goals and priorities. User-based evaluation enhances understanding of user satisfaction and guides improvements in design and usability, while task-based evaluation evaluates functional performance and identifies areas for enhancement.

Data Collection

After deciding on the research design, the subsequent step in evaluating chatbot performance is data collection. In a user-based evaluation, data can be gathered through surveys or questionnaires that capture users' feedback and perceptions regarding the chatbot's performance. These instruments allow users to express their satisfaction levels, opinions, and overall experience with the chatbot. On the other hand, in a task-based evaluation, data can be collected by assigning specific tasks to users and recording their interactions with the chatbot during task execution. This can involve monitoring the completion rate, accuracy of responses, response time, and any challenges encountered by users while performing the tasks. Data collection in both types of evaluation provides valuable insights into the chatbot's performance from different perspectives: user satisfaction and task accomplishment. By employing appropriate data collection methods, researchers can gather the necessary information to evaluate and analyze the chatbot's effectiveness in meeting user needs and successfully performing designated tasks.

Evaluation Metrics

When evaluating chatbot performance, selecting suitable evaluation metrics is essential. In a user-based evaluation, metrics such as user satisfaction, ease of use, and perceived helpfulness are commonly employed. User satisfaction measures the extent to which users are content with the chatbot's overall performance and their experience interacting with it. Ease of use assesses how easily users can navigate and interact with the chatbot's interface. Perceived helpfulness measures users' perception of the chatbot's ability to provide valuable assistance and address their queries effectively.In a task-based evaluation, different metrics come into play. Accuracy gauges the correctness of the chatbot's responses and its ability to provide accurate information or solutions. Speed refers to the efficiency of the chatbot in responding to user queries promptly. Completion rate measures the percentage of successfully completed tasks by the chatbot within a given time frame, indicating its task accomplishment capability. By utilizing appropriate evaluation metrics, researchers can quantify and analyze the chatbot's performance from different perspectives. These metrics provide measurable indicators of the chatbot's effectiveness, both in terms of user satisfaction and task performance, helping identify areas for improvement and guiding future development efforts.

Participants

Depending on your research design, you need to select participants.In a user-based evaluation, participants are chosen based on their demographic or psychographic characteristics to ensure a representative sample. This enables researchers to capture a diverse range of perspectives and experiences, enhancing the validity of the evaluation. On the other hand, in a task-based evaluation, participants are selected based on their proficiency and skills relevant to the designated tasks. This ensures that the evaluation accurately reflects the chatbot's performance in handling specific tasks and provides insights into its task accomplishment capabilities. By carefully selecting participants according to the requirements of the research design, researchers can gather relevant and meaningful data that aligns with the evaluation goals and objectives. This allows for a more accurate and comprehensive assessment of the chatbot's performance in the intended context.

Implementation

Once data collection is complete, the subsequent step is the implementation of the chatbot. This phase involves deploying the chatbot on a suitable platform, which can include popular platforms like Facebook Messenger, Slack, or embedding it within a website. The choice of platform depends on various factors, such as the target audience and the specific context in which the chatbot will be evaluated. For instance, if the aim is to reach a broad user base, platforms like Facebook Messenger may be preferred. Alternatively, if the evaluation is focused on a specific industry or organization, embedding the chatbot within their website or using a platform like Slack might be more appropriate. Implementing the chatbot on the chosen platform enables users to interact with it in a real-world or simulated

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

environment, depending on the evaluation setup. It allows researchers to observe and analyze how the chatbot performs in actual usage scenarios, providing valuable insights into its functionality, usability, and user satisfaction. Proper implementation ensures that the chatbot is readily accessible to the intended audience, facilitating a comprehensive evaluation of its performance.

Data Analysis

Finally, the collected data needs to be analyzed. For user-based evaluations, descriptive statistics can be employed to summarize and interpret the data, providing insights into user satisfaction, ease of use, and perceived helpfulness. This statistical analysis helps identify patterns and trends in the users' feedback and experiences. On the other hand, task-based evaluations may utilize inferential statistics to determine the significance of the chatbot's performance on specific tasks, such as accuracy, speed, or completion rate. Additionally, qualitative methods like sentiment analysis or thematic analysis can be employed to gain a deeper understanding of users' perceptions, attitudes, and emotions towards the chatbot. By combining quantitative and qualitative analysis techniques, researchers can derive comprehensive and meaningful insights from the collected data, allowing for a robust evaluation of the chatbot's performance.

Limitations

It is important to address the limitations of the study, such as sample size and participant bias, and their potential impact on the results. A small sample size may limit the generalizability of the findings, as it may not accurately represent the broader population. Participant bias, where individuals' opinions or behaviors are influenced by personal factors, could introduce skewed results. For example, participants with a specific background or prior knowledge may have different expectations or interactions with the chatbot. Recognizing these limitations helps contextualize the findings and informs the interpretation of the results. It is crucial to acknowledge and transparently discuss these limitations to provide a comprehensive understanding of the study's scope and potential implications.

Ethical Considerations

During the study, it is essential to address ethical considerations to ensure the well-being and rights of the participants. Obtaining informed consent is crucial, ensuring that participants are fully aware of the study's purpose, procedures, potential risks, and their right to withdraw at any time. Protecting participant confidentiality is another ethical concern, where measures should be in place to safeguard their personal information and ensure anonymity in reporting results. Additionally, ensuring the safety and well-being of participants is paramount. This includes minimizing any potential harm or distress caused by the chatbot interactions and providing necessary support or resources if needed. Adhering to ethical guidelines promotes trust, respect, and ethical conduct in research, fostering a responsible and ethical approach towards participant involvement and data handling.

IV. RESULTS

We conducted a study where we did interaction with different chatbot models from different fields using both domain specific and general daily based questions which consisted of some simple text chatbots, and some AI avatar based chatbots. We collected psychophysiological information from the participants while they were conversing with the chatbot, and at the conclusion of the experiment, we asked them to answer a few questionnaires regarding human-chatbot interaction, the realness of the chatbot, and its resemblance to a human.

The results that mattered for this study showed that users which were just domain oriented were happier and more satisfied while interacting with the simple text chatbot because it gave them the exact quick output they needed in completing their day to day task, but users which were not domain oriented and wanted a friendly chat with the chatbot were inclined towards the AI Avatar chatbot (In our case Replika), this was mainly because it was giving them a trustworthy conversation and close to human-like conversation as compared to simple text chatbot.

Replika AI's chatbot stores the user's personal information at the expense of providing them with pertinent output or dialogue in return.. The simple text chatbot was struggling in having an effective conversation when it came to knowledge outside of its domain.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

Since Human testing is time consuming and cannot be scaled, Automated testing of chatbots is the best way we can evaluate the chatbot with all the resources available on the internet and can be scaled as we want, we studied 2 ways of evaluating the chatbots which were BLEU and F-score. BLEU refers to Bi-Lingual Evaluation Understudy, which is a metric for automatically machine translated text. The BLEU score is a number between 0 to 1 evaluating that measures the similarity of the machine -translated text to a set of high-quality reference translations. On the other hand, the F-score is frequently employed for assessing information retrieval systems, such as search engines, as well as numerous categories of machine learning models, particularly in natural language processing.

When it came to the chatbot's execution, we experimented with several algorithms. We tried out the most popular ones that were widely available online and had a strong track record for creating chatbots. RNN (Recurrent Neural Network), CNN (Convolutional Neural Network), LSTM (LongShort Term Memory), and GRU(Gated Recurrent Units) were all used in this step of the procedure where LSTM and GRU are part of RNN. We used a drill down approach where we studied which algorithms are best suited for chatbots and we found out that RNN surpasses CNN when it comes to temporal data. Here, we learnt about the deep learning gate concepts and how they affect the processing of input and output. For instance, LSTM has three gates (Input gate, Forget gate, and Output gate), but GRU only uses two gates (Update gate, Reset gate). The output gate in an LSTM will take current input, the previous short term memory, and newly computed long term memory to produce new short term memory. The input gate in LSTM decides which information should be stored in long term memory, the forget gate decides which information from long term memory be kept or discarded and the output gate will take current input, the previous short term memory and newly computed long term memory to produce new short term memory. On the other hand GRU's update gate is responsible for determining the amount of previous information that needs to be passed along the next state and the reset gate is used to decide how much of the past information is needed to neglect it checks whether previous cell state is important or not. We know when it comes to conversational agents keeping a track of conversation is important as the user might relate its current query to some of the query it asked the conversational agent in the past and that information is handled using the concept of gates where we see that GRUs are easier to train and faster to run than LSTMs, but they may not be suitable for storing and accessing long term dependencies. We see that there is no "one" best type of RNN but they both serve different purposes in making different types of Conversational agents.

TABLE 1: Overview of the ratings of each functionality of the commercial Chabot's on a scale of 1 (very poor) to 5 (very good)

Chatbot Name	Visual Look	Form of implementation on the website	Speech synthesis unit	Presentation of knowledge and additional functionalities	Conversational Abilities, Language skills and context sensitivity	Personality Traits	Emergency responses in case of unexpected situations	Avg	Rating	Overall Quality
REPLIKA	5	4	4	4	6	4	3	4.125	4	Good
Nia (MIT ADT	3.5	4	0	3.5	3	3	0	2.6	3	Average
Quinn (Barclays)	3.5	3.5	0	2	3	3	0	2.31	3	Average
Naukri.com	4	4	0	4	4	4.5	3.5	3.5	4	Good
EVA (HDFC)	3	4	2	5	5	3	2	3.625	4	Good
Domino's	3.5	4	0	4	4	2	0	2.625	2.5	Average
ChatGPT	3	5	5	5	5	4	4.5	4.562	5	Best
Dotie (Indigo Airways)	4	4	0	4	4	4	3	3.375	3.4	Average
Alfered (Oven Story)	3.5	3.5	0	2.5	3	3	0	2.25	2	Poor

Note:The ranking is based on the research papers that were read.

V. CONCLUSION

The research project aimed to investigate the relationship between chatbot dialogues and their performance, focusing on identifying parameters that can be automatically measured to evaluate chatbots. To achieve this goal, six new

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 9, May 2023

performance criteria were developed based on existing evaluation techniques. The study extensively examined various natural language processing approaches and their corresponding algorithms for measuring chatbot performance. The research project compared and contrasted different evaluation methods, primarily focusing on two main approaches: questionnaire-based and automatic metrics.The questionnaire-based approach involved experts or users reviewing the chatbot and providing feedback based on their subjective experiences. This approach distinguished between quality of service measures, which evaluate how well the chatbot performs in terms of providing the desired outcome, and quality of experience measures, which assess how satisfied users are with the chatbot's performance.

On the other hand, the automatic metrics approach relied on natural language analysis techniques to measure specific criteria such as authenticity, confidence, readability, and sentiment. By conducting a literature review, the research project established that natural language analysis can provide objective measures of chatbot performance. For example, authenticity measures the extent to which the chatbot sounds like a human, while confidence evaluates the chatbot's ability to provide accurate and confident responses. Readability measures how easily users can understand the chatbot's messages, while sentiment gauges the overall tone of the chatbot's responses.

Additionally, the study identified six fundamental reasons why users may choose not to use chatbots available on the internet. These reasons include issues related to user interface, trust factors, effective communication, potential misunderstandings of queries, insufficient dataset answers, and language barriers. By identifying these reasons, the research project shed light on important considerations for chatbot design and development.

In conclusion, the research project highlights the significance of evaluating chatbots based on multiple performance criteria and leveraging automatic metrics to provide a more objective assessment. By investigating the relationship between chatbot dialogues and performance, and exploring various evaluation methods, the study contributes to the advancement of chatbot development and provides insights into creating more effective and user-friendly chatbot systems..

REFERENCES

[1]. Karolina Kuligowska, Commercial Chatbot Performance Evaluation, Usability metrics and quality assurance, 2019

[2]. Vijayaraghavan V, Jack Brian Cooper and Rian Leevinson J, Algorithm Inspection for Chatbot Performance Evaluation, 2020

[3]. Ram G Athreya, Ricardo Usbeck and AxelCyrille Ngomo, Enhancing community interaction with data driven chatbot- DBpedia Chatbot, 2019

[4]. Aleksandra Przegalinska, Leon Ciechanowski, Anna Stroz, Peter Gloor, Grzegorz Mazurek, In bot we trust:A new methodology of chatbot performance measures, 2019

[5]. Guendalina Caldarini, Sardar Jaf and Kenneth McGarry, A Literature Survey of Recent Advances in Chatbots, 2018

[6]. Sophia Keyner, Vadim Savenkov, and Svitlana Vakulenko, Open Data Chatbot: Towards Linked Open Data Conversational Interfaces, 2020

[7]. J. Haun, H. Chavez, and A. Nazi, The Impact of Chatbots in Healthcare Processes: A Systematic Review, 2019

[8]. Kun Zhou1, Kai Zhang, Yu Wu , Shujie Liu, and Jingsong Yu Unsupervised Context Rewriting for Open Domain Conversation, 2019

[9]. Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, Petter Bae Brandtzaeg, A longitudinal study of human–chatbot relationships, 2018

[10]. Krishna Gondaliya, Sergey Butakov and Pavol Zavarsky, SLA as a mechanism to manage risks related to chatbot services, 2018

[11]. Bibek Behera Chappie, A SemiAutomatic Intelligent Chatbot, 2019

[12]. Antje Janssen, Davinia Rodríguez Cardona, Jens Passlick, Michael H. Breitner, How to Make chatbots productive– A user-oriented implementation framework, 2019