Universidad Politécnica
de Madrid

Escuela Técnica Superior de
Ingenieros Informáticos

Máster Universitario en Software Engineering

Trabajo Fin de Máster

USER SATISFACTION ON CHATBOTS SYSTEM

Autor: Ramita Wisutmitnakorn
Tutor(a) interno(a): Xavier Ferré Grau
Tutor(a) externo(a): Raija Halonen
Universidad externa: University of Oulu

Madrid, June 2020

Este Trabajo Fin de Máster se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Máster
Máster Universitario en Software Engineering
Título: USER SATISFACTION ON CHATBOTS SYSTEM
Thesis no.: EMSE-2020-15

June 2020

Autor: Ramita Wisutmitnakorn

Tutor: Xavier Ferré Grau
LENGUAJES Y SISTEMAS INFORMÁTICOS E INGENIERÍA DE SOFTWARE (N)
ETSI Informáticos
Universidad Politécnica de Madrid

Resumen

The purpose of this study was to identify how the research trend of user satisfaction of chatbots in customer services is presented. As the recent raise of attention in chatbots system, especially in commercial use, the user satisfaction should be considered.

The main research question was how is the topic of user satisfaction of chatbot system in customer service presented in the prior academic research? Therefore, it was distributed into 3 sub-questions; how did the amount of research change according to time? How and why did the researchers conduct the research? How did the existing literature evaluate the user satisfaction?

Systematic mapping study research methodology was applied in the study. This research methodology considered the prior literature as primary studies then categorized them in order to get answer to research question. The results were how frequency the research had been published based on different scheme. 26 articles were involved as primary studies. The schemes included year of publication, research approach and user satisfaction evaluation approach.

The main contribution of this study was to discover trend regarding user satisfaction of chatbot in customer service context. This would help structure academic research area and motivate future research as well as being a guidance for conducting new research. The studied proved that the topic was still received a lot of interest from researchers as number of literatures regarding topic were growing. However, the results also stated that there was still lacking of research in some area or in specific scheme.

As this study's goal is to discover the research trend in order to identify the research gap, the future research is encouraged regarding the gap identified as well as the improvement of this study.

Abstract

The purpose of this study was to identify how the research trend of user satisfaction of chatbots in customer services is presented. As the recent raise of attention in chatbots system, especially in commercial use, the user satisfaction should be considered.

The main research question was how is the topic of user satisfaction of chatbot system in customer service presented in the prior academic research? Therefore, it was distributed into 3 sub-questions; how did the amount of research change according to time? How and why did the researchers conduct the research? How did the existing literature evaluate the user satisfaction?

Systematic mapping study research methodology was applied in the study. This research methodology considered the prior literature as primary studies then categorized them in order to get answer to research question. The results were how frequency the research had been published based on different scheme. 26 articles were involved as primary studies. The schemes included year of publication, research approach and user satisfaction evaluation approach.

The main contribution of this study was to discover trend regarding user satisfaction of chatbot in customer service context. This would help structure academic research area and motivate future research as well as being a guidance for conducting new research. The studied proved that the topic was still received a lot of interest from researchers as number of literatures regarding topic were growing. However, the results also stated that there was still lacking of research in some area or in specific scheme.

As this study's goal is to discover the research trend in order to identify the research gap, the future research is encouraged regarding the gap identified as well as the improvement of this study.

Tabla de contenidos

1. Introduction

The purpose of this study was to provide insight information on how the academic had accomplished the topic of the chatbots. As the trend of the chatbots grows rapidly, especially in commercial use, user satisfaction should be prioritized (Ren, Castro, Juan de Lara, 2019). Despite conducting a new experimental study, this thesis focused on the existing literature and categorized them.

Regarding the growth of online business and the number of customers, customer service also faces some difficulties in serving all needs for each one. There are various approaches for customers to reach the company and ask any query; phone call service, e-mail service, live chat service, etc. Even though the many channels of service are provided, the customers still need to wait for a long time as human resources could be limited. Moreover, the customers are possibly got wrong information from human errors (Thomas, 2016). In order to solve that problem with human-based, automatic services have been developed. Chatbots might be the best well-known one.

The chatbot is a computer program that receives massive attention. The reason is that it is more cost-effective, time-saving comparing to humans. As per serving the need in the industry, the new chatbot studies have been intensively published in these recent years in order to seek a better solution (Ren et al., 2019). Like any other research topic, the gap between academic publishes and the real-world industry could exist. Consequently, the current academic position needs to be discovered. This thesis can be the starting point of another research.

The main research question was developed regarding the trend of academic research. Therefore, it is set as followed.

> How is the topic of user satisfaction of chatbot system in customer service presented in the prior academic research?

To answer the research question, it was distributed into sub-questions:

> Q1: How did the amount of research change according to time?
>
> Q2: How and why did the researchers conduct the research?
>
> Q3: How did the existing literature evaluate the user satisfaction of chatbot service?

The purpose of Q1 was to look for the research trend over time and to see whether it was still in the middle of interest among researchers. Q2 was meant to find out the environment setting in research, for instance, in a controlled environment or from the real-life production. Moreover, it was meant to find out the objective of the research. The last question, Q3, was to specify the method used for the evaluation of user satisfaction.

The research method using in this thesis was a systematic mapping study. It could help generate a fresh idea of academic research. Moreover, it could help identify the gap between academic research and the industry (Peterson, Vakkalanka & Kuzniarz, 2015). The primary studies which were included in this

thesis were all experimental. The theoretical ones were excluded as they could not help answer the research questions.

The main contribution of this study illustrated the trend of the research area of chatbot service in the context of user satisfaction. It provided the insight information of academic articles in classification pattern. An outcome as well as a discussion could be a motivation for new research in the future.

The structure of the report follows the guidance provided by the University of Oulu. In chapter 2, the prior literature about chatbot definition and the chatbot service in the context of social media are presented. The chatbot evaluation prior to previous studies are also introduced as well as the similar and related existing research. Chapter 3 describes the research method, systematic mapping study. The details and procedure of conducting the research based on published articles are presented. Chapter 4 provides the details of implementation based on chosen research methodology. The following chapter 5 presents the findings according to research questions while chapter 6 is a brief discussion from the result of the thesis. Lastly, chapter 7 contains the conclusion of this thesis.

2. Prior Research

This chapter presents prior studies related to the topic. In chapter 2.1, chatbots system definition and context of use are presented. Chapter 2.2 consists of chatbots evaluation metrics as it presented in previous published articles. Lastly, chapter 2.3 looks insight into the similar research considering their research procedure and results.

2.1 Chatbots system

Shawer & Atwell (2007) give the definition of the chatbot as a computer program that interacts with the users. It is applied the natural language processing (NLP) in order to be able to hold the conversation with humans. The chatbot is widely used in various ways, for instance, it can be used for entertainment, education, information retrieval, especially in e-commerce, which is the topic for this thesis.

The technology-based of chatbot system is artificial intelligence. Key strengths of chatbot include a user-friendly, conversational, knowledgeable and quick response. Developers should design the mood and pattern of it upon the objective and target users. It can hold a smooth conversation with the user if it knows how to answer those questions. Therefore, developers must train the system and compute data as much as possible considering the scope of the system (Dahiya, 2017).

Chatbot system is simple. Figure 1 illustrates overall interaction of chatbot system. Firstly, the system gets the text input from user, it will compare with its database in order to find the proper response. Once the system can find proper text output, it will send that to the middle device. Therefore, the most challenging part is to define the accurate output (Dahiya, 2017).

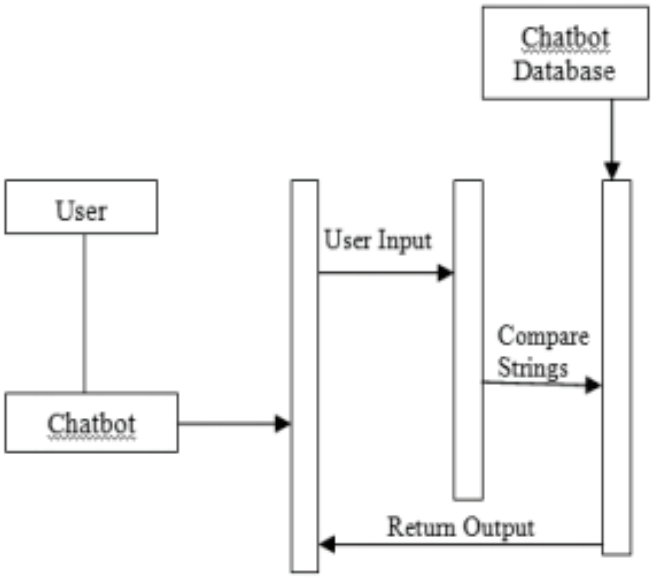

Figure 1. Sequence diagram of chatbot design (Dahiya, 2017, p.159).

A new challenge for software engineers is to develop proper and smart solutions for chatbots. The most common approach is an information retrieval and template rules which are also known as a deep learning technique. The steps of implementation, presented by reference article, include cleaning the irrelevant data, building the vocabulary, adopting word embedding features and training the network. This is also an approach for any deep learning network (Xu et al., 2017).

There are similar terms to chatbots and that always make people confused between them. The most well-known one might be conversational agents. Conversational agents are the combination of verbal and non-verbal human-computer interaction system. However, sometimes in academic research, terms can be used in place of another (Kopp & Wachsmuth, 2004; Io & Lee, 2017).

Chatbot is now taking over human's jobs. As an entertainment tool, the chatbot had been developed to simulate human conversation as fictional or real person characteristics. It is the first intention of implementing the chatbot according to chatbot history Moreover, the system could be developed as a learning assistant. There are also multiples programs which meant to help language learning. Moreover, it can act as a middle-man between teachers and students. It helps teachers find out whether students have any problems as well as answer student queries. In the business part, chatbot plays an important role in e-commerce. It can be a shopping assistant which offer help in the store or provide information in the store. For online shopping, it can be the first customer service channel to reach because it can be available anytime (Shawar & Atwell, 2007).

Currently, there is the massive usage of online service. People always use social media sites to contact the company. It is usually a part of chatting applications or websites. As traditional service of using humans is time-consuming and able to make mistakes anytime. The study also shows that human-based takes a longer time to reach customers than customers expect. In order to improve their customer service, many companies would like to invest in chatbot technology (Xu, Liu, Guo, Sinha & Akkiraju, 2017; Peras, 2018.)

2.2 Chatbot evaluation metrics

The assessment of chatbot is said to be a challenging task as there is no clear explicit measurement. It is hard to define how well the system works or make the comparison between different system. However, there are many studies regarding the improvement of chatbot in technology perspective but not so many are based on other's point of view, for instance, from business or user. Therefore, chatbot evaluation in the context of user motivation is still a difficult task as there is no explicit metric defined because of lack of interest. Despite the lack of primary studies, the writer emphasizes chatbot evaluation and set it into 5 perspectives; user experience perspective, information retrieval perspective, linguistic perspective, technology perspective and business perspective (Hung, Elvir, Gonzalez & DeMara, 2009; Paras, 2018). All perspectives and their categories are presented in table 1.

Table 1. Chatbot evaluation perspectives and categories.

Perspective	**Category**
User experience perspective	Usability Performance Affect Satisfaction
Information retrieval perspective	Accuracy Accessibility Efficiency
Linguistic perspective	Quality Quantity Relation Manner Grammatical accuracy
Technology perspective	Humanity
Business perspective	Business value

For user experience perspective, the metrics are mostly quantitative. This type of perspective contains four categories; usability, performance, affect and satisfaction. Usability means efficiency and effectiveness. Performance is the completion of tasks upon user's goals. Affect evaluates experience and emotion of the user. Satisfaction is pleasure of user upon their expectation of the system. Both affects and satisfaction are hard to measure and define because it heavily depends on individual's experience and goal. Sample metrics consists of a rating scale, surveys, questionnaires, support of Help and Cancel command, number of responses from user (Paras, 2018).

The second perspective, information retrieval, represents the aim of getting accurate and relevant required information from the system. This has three categories; accuracy, accessibility and efficiency. The metrics are qualitative. Sample metrics consist of turn correction ratio, total number of turns per tasks, typing error and synonym, precision, recall (Paras, 2018).

Linguistic is the third perspective. This one evaluates the rate of linguistic accuracy and rate of returning correct response with the suitable vocabulary and grammar. The categories of this perspective are quality, quantity, relation, manner and grammatical accuracy. The metrics belonging to this group are qualitative, for instance, total number of errors, vocabulary range, spelling check, grammar check (Paras, 2018).

Technology perspective evaluates how well the system acts like human. Therefore, there is only one category; humanity. It checks the attribute of naturalness. The well-known metric belongs to this is Turing test. Moreover,

rating scale, percentage of success and rejection can also be considered (Paras, 2018).

Business perspective is the last one. It assesses the business value of the system. Business value refers to effectiveness against cost. Metrics of effectiveness consists of number of users, duration of conversation and number of conversation while cost's metrics are number of agents in conversation, number of unsuccessful conversations, number of repeated queries (Paras, 2018).

Shawar & Atwell, 2007 propose the evaluation metrics which used to assess chatbots. In the study, 3 evaluation metrics are presented; dialogue efficiency, dialogue quality metrics and user satisfaction. The first metric, dialogue efficiency, measures how accurate the system can response to user's queries. Secondly, dialogue quality metric, measures reasonableness of the system. It categorizes the response into 3 categories; reasonable reply, weird but understandable, nonsensical reply. Last metrics is user satisfaction which gather the data from direct user feedback. At the end, it is suggested that the chatbots system evaluation should not adopt established approach. It is better to adopt one customized for the individual system.

Hung et al., 2009 describe the established metrics which can be used for chatbots software evaluation. The metrics system is called the PARAdigm for Dialogue System Evaluation (PARADISE). User satisfaction is the main purpose of this system. Task success and dialog cost are considered in this context. Task success means the completion of system response towards user goals. Dialog cost can be used for 2 aspects; efficiency and quality. Efficiency is the resource used to complete the task while quality measures chat log's content.

2.3 Related mapping study research

Io & Lee, 2017 present the study of the prior literature about chatbot using bibliometric analysis. Bibliometric analysis is a methodology to explore the trends of a research topic in order to encourage and help researcher to conduct future research. 583 literature are included in the study. They are clustered base on different perspective.

The result from the study indicates that chatbot topic becoming sudden popular in 2015 as artificial intelligence and related technologies have been well-developed and came to light among researchers. Another result is that technology used in chatbot has been changed. Classical Natural Language Processing (NLP) is replaced by deep learning. Moreover, the major studies are now mostly in education field. However, there is also great possibility that it can be used in the other field, for instance, business. Last finding mentions that most of studies are focused on technical perspective and lacking of human and business point of view perspective (Io & Lee, 2017).

Various suggestions for future research are also proposed. According to the findings, new technology should be paid attention as it is now in time of technology transition. New platform like mobile applications should also be considered since smartphone is truly popular. Lastly, new research in other perspective beside technology is encourage to be studied (Io & Lee, 2017).

Ren et al., 2019 present another systematic mapping study regarding chatbot. In the study, it is indicated that there were only few system mapping study research of this topic in the past. The research focuses on usability of chatbot system. The search string consists of two parts; usability and chatbots. In case of chatbots; several synonyms are used including conversational agent. Finally, total of 19 papers are included as primary studies in the research and clustered based on usability techniques, usability characteristics, research methods and types of chatbots (effectiveness, efficiency and satisfaction).

The result indicates that usability techniques are not well presented in primary studies. Moreover, they are used for evaluating the developed chatbot. Evaluation should be performed according to context. For research method, the most common ones are surveys, experimental studies and usability tests (Ren et al., 2019).

3. Research Method

In this chapter, the systematic mapping study (SMS) is described. The definition and comparison to the similar research methodology, Systematic Review (SR) are presented in Chapter 3.1. For second part of this chapter, the procedure of system mapping study is defined.

3.1 Research method selection

There are two similar approaches for summarizing the published articles and provides insight information based on those evidence. They are systematic review and systematic mapping study. Both research methods are similar and often make the researcher confused between them. The objectives of these two are different even though the approaches are quite similar (Denyer and Tranfield, 2009; Peterson, Feldt, Mujtaba & Mattsson, 2008; Peterson et al., 2015).

Systematic review research method is derived from medical research. Its objective is to evaluate the evidence and generalize them in order to avoid the same errors as in prior research. There is a well-define steps and procedure provided. The research questions contain specific goals towards a specific topic. Therefore, in one topic area, research can be developed from a different perspective. Besides that, quality assessment is an important part. The advantages of research method consist of reducing bias, more general conclusions from wide range data. In contrast, the disadvantage of it is effort consuming as there are many steps required to conduct this type of research (Denyer and Tranfield, 2009; Peterson et al., 2008; Peterson et al., 2015).

Systematic mapping study is for defining the research area via classification. The research questions aim to discover the trend of the current study and research structure then provide the overview summary. The main difference is the research progress. The search is heavily based on the research question and focuses on one specific topic. Moreover, the quality of the assessment is not required for this type of research. It can be said that a systematic mapping study could be the initial part that inspires systematic review research. The outcome of this is the overview of the presented research form on classification (Peterson et al., 2008; Peterson et al., 2015;).

Both research methods are different in many perspectives. Firstly, the goals are completely unalike. Systematic mapping study provides state of research area whereas systematic review provides the best practices according to existing literature. Moreover, in mapping study, more articles are involved but do not go deeper in details. On the other hand, systematic review considers the details of each primary studies. Therefore, systematic review requires more effort and time than systematic mapping study. Lastly, because systematic review specifies the topic area and sometimes the method, the number of relevant research decreases. Therefore, bias can be easily introduced (Denyer and Tranfield, 2009; Peterson et al., 2008). In conclusion, systematic mapping study was selected in this study, as it could obviously answer research questions.

3.2 Systematic mapping study procedure

Systematic mapping study is known as the common research method in the software engineering field. The objective is to understand the current state of art of research based on published articles and help in conducting reliable studies in the future. The overall amount of this kind or research is growing as there are more than enough research to act as primary studies. It is used to define the classification and structure a research area. The results are based on the amount of publications as presented and categorized in different scheme (Peterson et al., 2008, Peterson et al., 2015; Ren et al., 2019).

Peterson et al., 2008 establishes the guideline of systematic mapping studies in software engineering. Afterwards, Peterson et al., 2015 then updates the guideline. The steps inherited from both papers include defining research questions, conducting the search for primary studies, selecting the relevant papers, extracting data and lastly, analysis and classification. The overall procedure is presented in figure X.

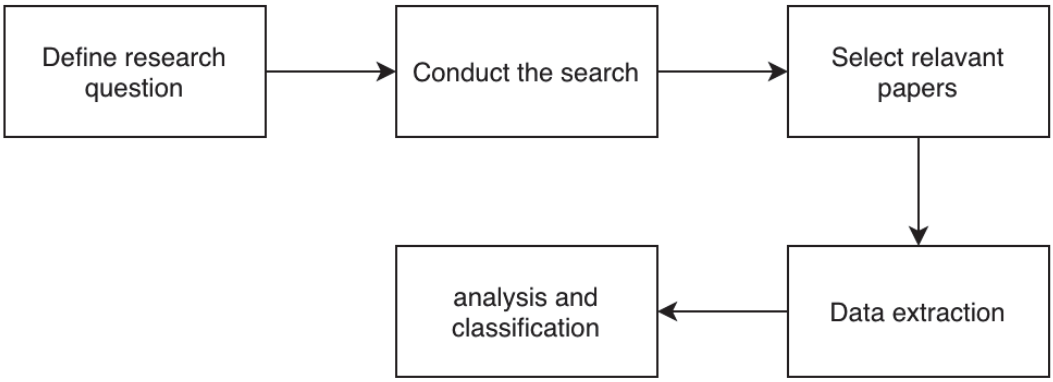

Figure 2. Systematic mapping study procedure.

As shown in figure 2, the first step is defining research question. It helps to identify and scope to match the research goal. Mostly, research questions are focused on coverage of specific topic. Questions about venues, research methods and trend are commonly used (Peterson et al., 2015).

Second step is conducting the search. Search strategy is chosen first. There are snowballing, manual search and database search. Database search is the most popular one while manual search is more effective comparing to other two strategies. It is possible to do multiple strategies as well but it is said to be time-consuming. The search strings are identified based on research topic. If the search result is in large number, it is better to restrict the search strings. After gathering relevant papers from search result, evaluation is needed. It can be performed by another expert (Peterson et al., 2015).

Third step is screening the papers. Usually, pre-set criteria are identified beforehand. It may refer to relevance of the papers, time period or language used

in papers. In order to improve the screening process, strategies should be proposed. First is to have protocol which is reviewed by multiple researchers. After that, the selection is done on title and abstracts of the papers. If there are multiple researchers, each can work individually then have discussion later. The second round of screening is to real the whole paper and identify whether it can be included as primary studies (Peterson et al., 2008; Peterson et al., 2015).

The next step is data extraction. As final primary studies have been selected, the information is gathered. There are two options to perform this task. First it to have more than one researcher. The first person performs data extraction while another also performs individually then the result is discussed. Alternatively, second person can act as reviewer. Another option is to have assessment over criteria based on sample set of articles. The first option, having two or more researchers, is more preferable (Peterson et al., 2015).

Fifth step is analysis and classification. The information is categorized according to the topic or research question. The papers are sorted according to scheme. Excel can be utilized as it can be easy for reviewers. The analysis is how often the papers published considering different theme. This mapping can help researcher generate the idea of current research status as well as identify the gap. The mapping is presented as a result of research. Any visualizations are highly recommended, for instance, bubble plots, bar plots, pie chart. (Peterson et al., 2008; Peterson et al., 2015).

The treats to validity of systematic mapping study are also presented. The results are heavily based on primary studies. Therefore, the treat is developed based on selection criteria. Other treats consist of publication bias (theoretical validity), poorly data extraction performance (descriptive validity), researcher bias (theoretical validity), quality of primary studies (theoretical validity) and reliability of the conclusions (Peterson et al., 2015).

4. Application of research method

In this chapter, the implementation of systematic mapping study research method is presented. The actual activities done in the study are mentioned in details. The sub-chapters follow the steps described in chapter 3.2.

4.1 Define research question

The main research question was

> How is the topic of user satisfaction of chatbots system in customer service presented in the prior academic research?

It tried to discover the research trend of chatbot quality from user perspective as the system had been used in context of customer service. To be more detailed, it was set into three sub-questions.

> Q1: How did the amount of research change according to time?

The answer to this question was the amount of primary studies categorized by year of publish. The objective was to determine whether this topic is still in the middle of researcher's attention.

> Q2: How and why did the researchers conduct the research?

This was meant to find out the research approach using in primary studies. The research approach and objective of the research were discovered to answer this research question.

> Q3: How did the existing literature evaluate the user satisfaction of chatbot service?

As there were various ways to assess the satisfaction of users, this sub-question tried to identify and categorize the user satisfaction evaluating activities as well as the metrics used.

4.2 Search for the primary studies

Second step was searching for primary studies to be data of the study. The appropriate way to define search strings is to isolate the key index into sets. After that, the term from each set is combined. The published year was limited to 2010 – 2020. The result from this step is the published articles gathering from all search strings.

The data source was Google Scholar. Google Scholar was a search engine for scholarly literature. Access grant was provided by University of Oulu. Therefore, only available articles from university's permit were included in the study.

As the primary studies were required as data for the thesis. First of all, the sets of search strings were defined.

- Set1: Scoping the search to be only software engineering. In this thesis, 'software engineering' was used.
- Set2: Search directly related terms. In this thesis, these were divided into two lists:

Set2.1: ‘chatbot’ ‘conversational agent’
Set2.2: ‘user satisfaction’ ‘customer satisfaction’.
    - Set2.3: term regarding to customer service; ‘customer service’
- Set3: Search the terms related to. In this thesis, ‘case study’ ‘experimental study’ were used

After the search strings of each set had been defined, they were combined and used in Google Scholar. It was called first string revision. The sample of combination included ‘Software engineering chatbot user satisfaction case study’, ‘Software Engineering chatbot customer satisfaction experimental study’. However, the relevant articles were not enough. Search strings had to be revised. For second revision, search string was more general. The search combinations were shortened and ‘software engineering’ term was completely removed because it was too specific. Moreover, some sub-sets of set 1 were not included in some combinations.

- Set1: Search direct related terms.
    - Set1.1: terms regarding chatbot system; ‘chatbot’ ‘conversational agent’
    - Set1.2: term regarding user satisfaction; ‘satisfaction’
    - Set1.3: term regarding to customer service; ‘customer service’
- Set2: Search the terms related to. In this study, ‘case study’ ‘experimental study’ were used.

Table 2. Second revision of Search string combinations.

String ID	Set 1.1	Set 1.2	Set 1.3	Set 2
ST1	chatbot	satisfaction	customer service	case study
ST2	chatbot	satisfaction	customer service	experimental study
ST3	conversational agent	satisfaction	customer service	case study
ST4	conversational agent	satisfaction	customer service	experimental study

Table 2 presents the search string and its assigned ID. This string ID was used in the rest of the report. The search string was input in Google Scholar search engine. All sets were combined and connected by space. For example, ST1 search string was ‘chatbot satisfaction customer service case study’.

4.3 Study selection

Study selection was performed twice. In first round, due to great number of articles gathered, only first 50 relevant articles from each search string were selected. They were assessed against pre-set criteria. The paper would be excluded if

- It was not written in English.

- It was not accessible with university's grant.
- It was duplicated with other studies.
- It did not include case study
- It was not in software engineering field
- It was relevant with the topic research; chatbot in customer service

Thanks to feature provided by Google Scholar site, there was a setting to select literature in specific language. Hence, only articles written in English were in the search result. If it could not be accessed then there was no way to assess it. It would be unfortunately eliminated. Moreover, if it also appeared in previous search string's result, it was considered as duplicated and removed from the latter list. Lastly, by reading its abstract, when the paper was not in software engineering field or not relevant with the study topic, it was excluded as well. It was considered as irrelevant if it was not case study nor did not help answer the research questions.

After getting rid of some irrelevant papers, the remaining were performed second-round selection. At this step, the entire paper was read. Once again, if it was not able to answer the research question, it would be eliminated. The number of papers at each step is provided according to each search string in Table 3.

Table 3. Number of results from each round selection.

String ID	Initial results	Results after first-round selection	Results after second-round selection
ST1	4610	28	17
ST2	2450	13	3
ST3	17000	8	4
ST4	17300	5	2
Total	41360	54	26

Table 3 presents the number of articles gathered at each step per each search string. At first-round selection, only first 50 relevant articles according to Google Scholar ranking algorithm were assessed for each search string. Most literature from ST1 passed pre-set criteria because the search string was similar to research topic. ST2 was also similar but many were removed because of duplication. ST3 and ST 4 contained more board term, 'conversational agent'. Results from both consisted of other technology than chatbot, for instance, call agent or voice-controlled robot. Therefore, most of them were eliminated. In the end, 26 literature were selected as the primary studies and proceeded to the next step. The final selected papers were listed in Appendix A.

4.4 Data extraction and classification

Information in each literature were extracted and categorized. Extracted information were relied on research question and meant to answer them. For

better understanding, the data were extracted as in following table. Spreadsheet was utilized to analyze those data. After the data extraction had been done, all information was analyzed. Each paper was categorized and counted in different aspects. The result of categorization as well as its details were presented in Chapter 5.

Table 4. Data extraction fields.

Data	Description	Related Research Question
Paper ID	'PA' + Unique number	
Article Title	Name of paper	
Author Name	Set of author's names	
Year of publication	The year that article had been published	Q1
Research environment	Specifying in which environment the research had been conducted	Q2
Research objective	Purpose of conducting research	Q2
Satisfaction evaluation methods	Practice used to evaluate user satisfaction	Q3
Satisfaction evaluation perspective	Metric applied for user satisfaction	Q3

Table 4 illustrates the different information as result from data extraction. All papers were sorted by the title. Paper ID started 'PA' and followed unique integer number assigned to each article. Paper ID was used as a reference in the entire report. Article Title was the name of paper. Author Name represented the name of every author. Year of publication was the time when the article had been published. This data helped answer research question Q1. Research environment specified how the researcher set up the research. Research objective meant the purpose of study. Both research environment and research objective were related to research question Q2. Last two fields, user satisfaction evaluation methods and user satisfaction metrics, were used to discover how the researcher evaluate the system and meant to answer the research question Q3.

5. Results

This chapter presents the result of this study which was inherited from data extraction and data analysis of 26 primary studies. Sub-chapters follow the order of research questions respectively. Chapter 4.1 presents the year of publication and research trend changed over time. Chapter 4.2 presents the finding of research approach applied in primary studies. Last chapter, 4.3, describes the evaluation method of user satisfaction. Therefore, those findings are discussed in the next chapter.

5.1 Year of publication

Research question 1 was to discover the change of research trend over the time. The amount of research published each year was presented in this section.

Q1: How did the amount of research change according to time?

The date range of search results was limited to be only papers published from 2010-2020 in order to focusing on current trend and getting rid of old research which might not be valid anymore. Therefore, the number of articles published in each year is presented in following table 5 and figure 3.

Table 5. Categorization based on year of publication.

Year of publication	Paper ID	Number of papers
2010	-	
2011	-	
2012	-	
2013	PA26	1
2014	-	
2015	-	
2016	PA1	1
2017	PA22	1
2018	PA6, PA7, PA8, PA9, PA12, PA13, PA19, PA20, PA21, PA25	10
2019	PA2, PA4, PA10, PA11, PA14, PA15, PA16, PA17, PA24	9
2020	PA3, PA5, PA18, PA23	4

Table 5 presents the papers categorized by year of publication. The paper ID and amount of papers are shown.

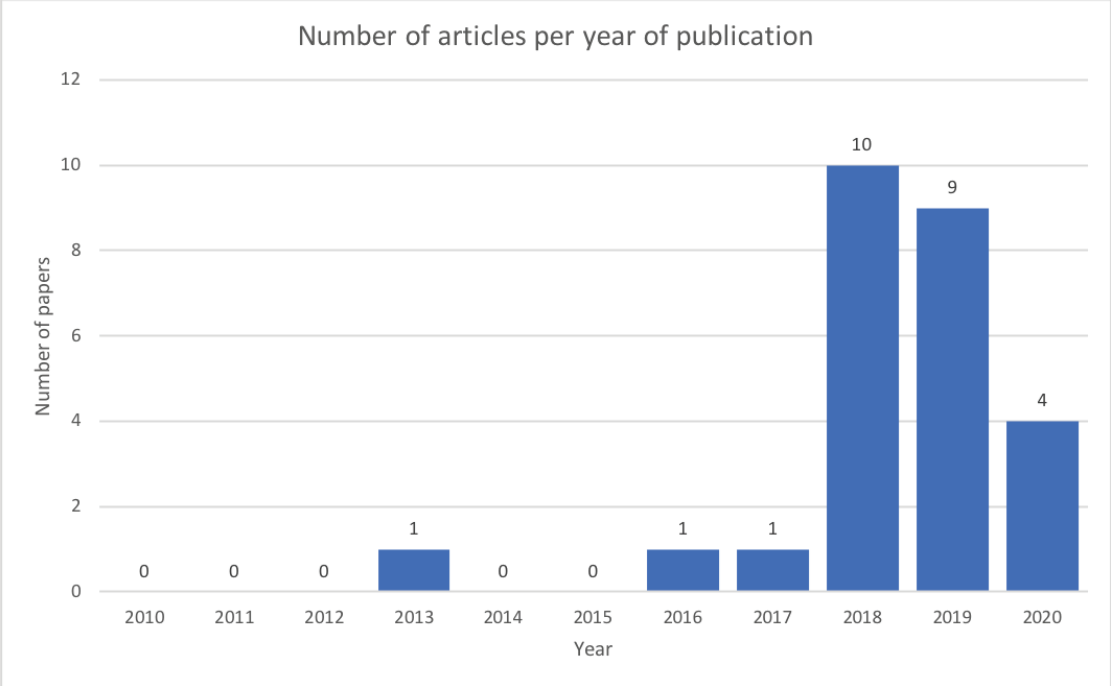

Figure 3. Frequency of research published according to year.

The figure 3 displays the same data in graph manner. As shown in table 5 and figure 3, the oldest article was published in 2013 despite the fact that the search was set to start from 2010. Moreover, there was only one paper in that year. The trend shows that more papers were published in more recent years. The increase of papers was presented starting from 2018. Even at the time that this study had been done (May, 2020), there were already 5 papers published this year.

The earliest paper PA26 was published in 2013. This paper presented new chatbot system implementation approach. The researcher also conducted the assessment of this system with target participants. Assessment was done in setup environment. Although, there was no paper published in next two years, the next paper, PA1, was done in 2016 following by PA22 in 2017. PA1 and PA22 studied also new chatbot systems and validated the requirements and system quality from user satisfaction perspective. PA1 proposed new system as mental healthcare advisor while PA22 focused on chatbot in e-commerce website.

In 2018, 10 papers were published which was the greatest amount from the search criteria. The topic was acknowledged by the researchers as more papers were published. More variety research methods were applied. PA8 focused on the real production and collected data from real users. The same trend still strongly remains in 2019-2020.

5.2 Research approach

The research question 2 focused on which environment the researcher conducted the research as well as the objective of study. The analysis was done in two different perspectives; environment of the research and its objective.

Q2: How and why did the researchers conduct the research?

Firstly, research environment in this study meant how researchers set up the research; in controlled environment or from real system and actual experienced users. This setup also affected the result of research. Result from categorization of the study is illustrated in table 6 and figured 4.

Table 6. Categorization based on research environment.

Environment	Paper ID	Number of papers
Experimental study	PA1, PA2, PA3, PA4, PA5, PA6, PA8, PA9, PA11, PA12, PA14, PA15, PA17, PA19, PA20, PA21, PA22, PA25, PA26	19
Case study	PA7, PA10, PA13, PA16, PA18, PA23, PA24	7

Table 6 presents the papers categorized by research environment. The paper ID and amount of papers per each category are shown.

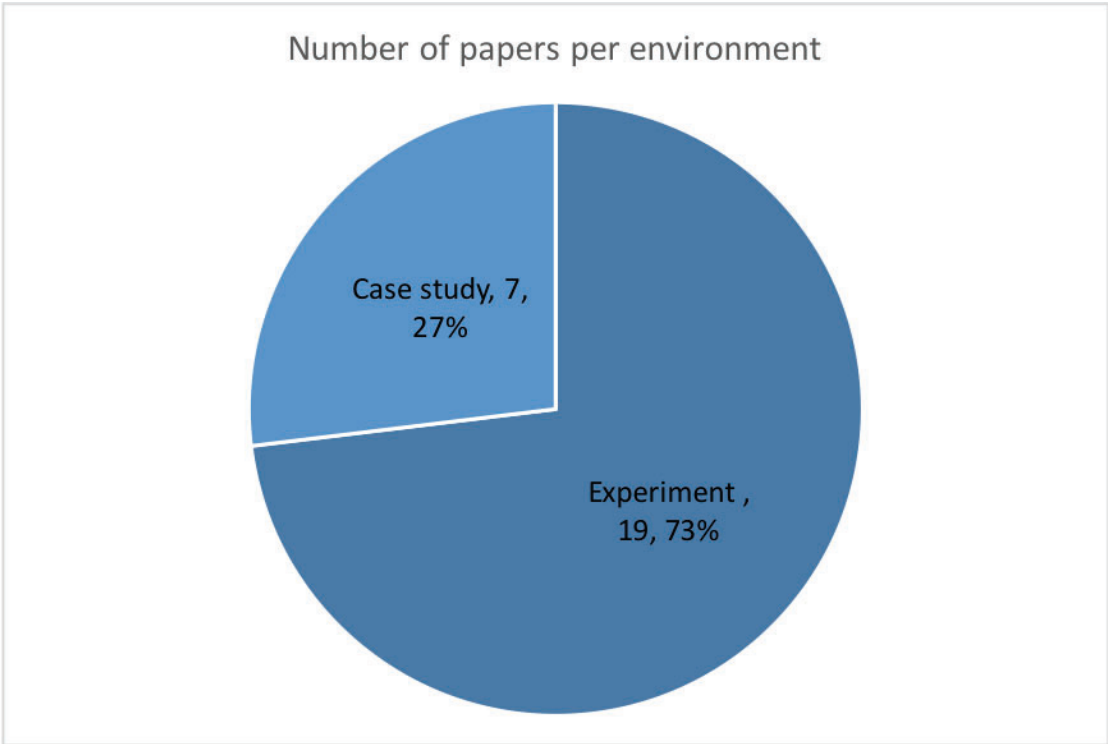

Figure 4. Frequency of research published according to research environment.

Figure 4 shows the number of papers categorized into two environments; experimental study and real production. The majority was experiment study as 73% of all papers were conducted in this environment whereas the rest, 27% of papers, were collecting from real system which used by real user without any control.

19 studies or 73% of primary studies were conducted in experiment environment which means the actions of users were monitored by researchers. The research was controlled fully or partially by researchers. In this kind of research, tasks were usually given to the selected research participants with some guidance. Users needed to complete them then provide the feedback in order to do further evaluate. Among all 19 experimental studies, there were some studies which dealing with new implementing system.

On the other hand, 7 out of 26 papers stated that their data were gathered from the real production and from the real user. There was no control or guidance provided to users. In real production environment, there were several ways to collect the data from user. Paper PA7 was the only one of all paper which conducted the research with developers while others papers did with users. PA7 interviewed the managers and developers of the product. PA10 used phone interviews method to collect the data. PA13 and PA24 did data analysis from the comments and chat history. PA16, PA18 and PA23 asked the users to participate in survey after using the system.

For second perspective, all of primary studies were studied about user satisfaction but based on different objective. Those objectives could be categorized into 3 categories; new system proposed, chatbot technology and algorithm improvement, existing system evaluation. New system proposed meant that the studies were conducted to serve new requirements in specific context. Normally, the system was proposed and case study was conducted to validate the usage. Another category, chatbot technology and algorithm improvement, was for studies which adopted different technology and technique and assess its effectiveness. Last one, existing system evaluation, meant for studies that evaluated the chatbots which were used in real situations. The summary of categorization is presented in following table 7 and figure 5.

Table 7. Categorization based on research objective

Category of objective	Paper ID	Number of papers
New system proposed	PA1, PA4, PA9, PA10, PA12, PA21, PA22, PA25, PA26	9
Technology and algorithm improvement	PA2, PA3, PA5, PA6, PA11, PA14, PA17, PA19, PA20	9
Existing system evaluation	PA7, PA8, PA13, PA15, PA16, PA18, PA23, PA24	8

Table 7 presents the papers categorized by research objective. The paper ID and amount of papers are shown.

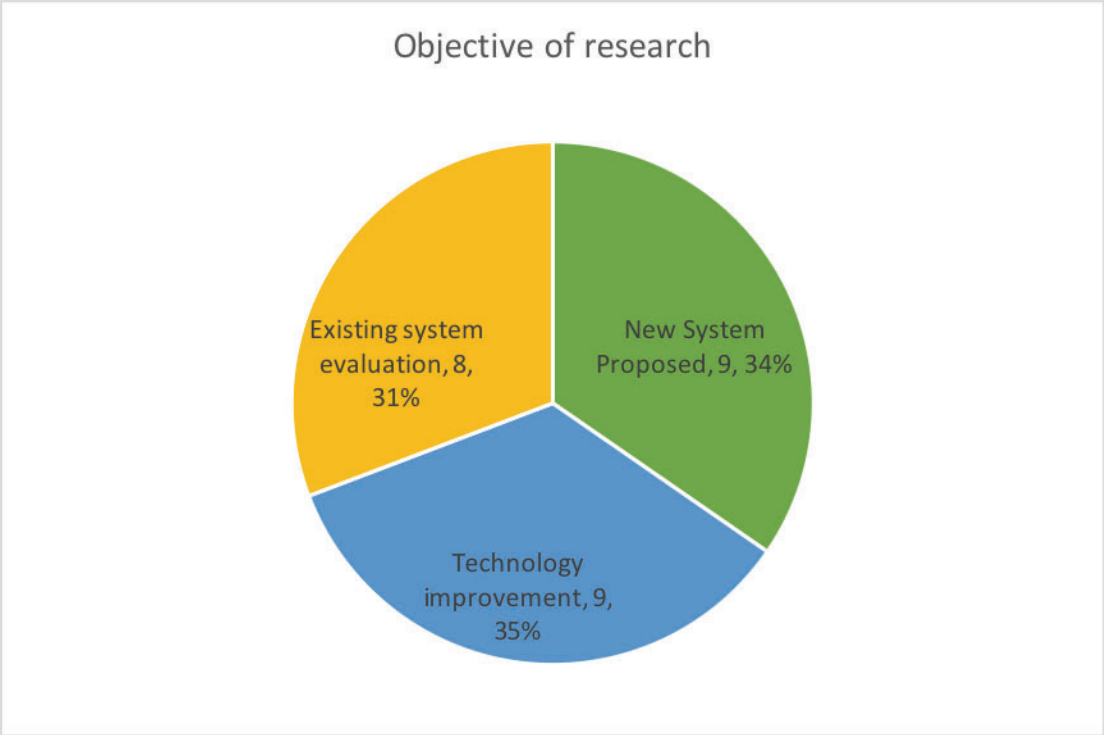

Figure 5. Frequency of research published according to objective of research.

Figure 5 displays the pie chart for objective of the primary studies. The categories consist of existing system evaluation, new system proposed and technology improvement.

According to Table 7 and Figure 5, there were 9 out of 26 papers which conducted for new system proposed. The research would propose the system for specific user and purpose. For example, PA1 proposed new system as a mental health advisor especially for drinking behavior. The study also included the case studied and evaluation from chosen participants. Another example was PA26. In this research, new chatbot system as an undergraduate advisor was presented. The system helped answer the questions regarding the university.

Secondly, there were 9 papers considering technology improvement. Various techniques and technology had been adopted in academic research. Even they were about technology and algorithm stuffs, the evaluation was based on user satisfaction. PA19 studied about how delay response affected the user experience and how user felt about delay response comparing to instant response.

Last category was existing system evaluation. Total of 8 research belonging to this category studied about positive and negative impact as well as business value of the chatbot system as it had been used as commercial product. PA7 presented succeed chatbot system in South Korea by interviewing developers considering business value of the system. PA23 was to find out the product value in terms of user satisfaction by a survey with users in Indonesia.

For better understanding, these two perspectives, environment and objective of research were analyzed together. Figure 6 presents the finding from the analysis.

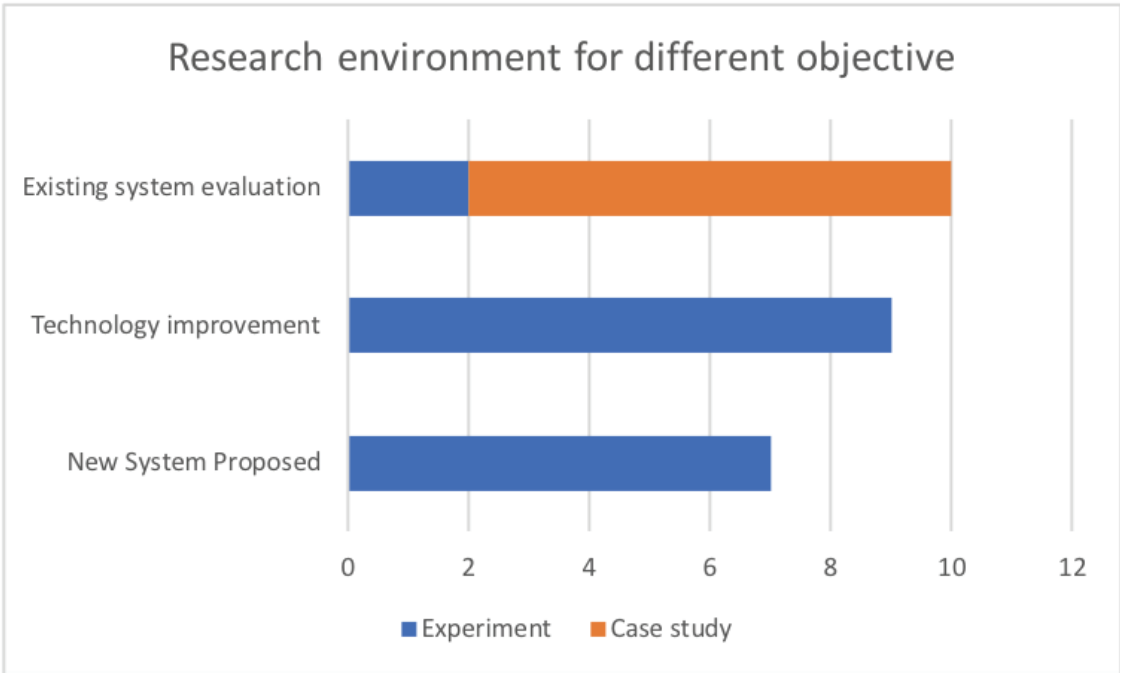

Figure 6. Frequency of research published according to research environment and objective.

Figure 6 presents the environment (experiment or case study) that was chosen for different objective. The result shows that when the research was for evaluating the existing system, the researcher adopted both controlled environment and field observation approaches. However, case study was still more preferable by the researchers.

On the other hand, for technical matters, more research adopted experiment study approach over case study. As there was only one paper chose case study approach, PA10. In this research, users were interviewed based on their experience with the real system. None of articles adopted case study approach. All were experimental study. For last objective, new system proposed, because of being a new system, it was impossible to do it in case study manner.

5.3 User satisfaction evaluation

Research question 3 was meant to discover how the user satisfactions had been evaluated and in which aspects they were assessed.

Q3: How did the existing literature evaluate the user satisfaction of chatbot service?

The findings report is divided into two parts, evaluation method and evaluation perspective, respectively.

5.3.1 Evaluation method

Many of the primary studies applied more than one method when it came to evaluation. Therefore, it might not be meaningful to calculate the percentage. The summary of evaluation methods is displayed in the table 8.

Table 8. Evaluation methods and papers categorization.

Evaluation method	Paper ID	Number of occurrences
Survey and questionnaire	PA1, PA3, P4, PA6, PA8, PA9, PA11, PA12, PA13, PA16, PA17, PA18, PA19, PA20, PA21, PA23, PA24, PA25, PA26	19
Interview	PA1, PA4, PA7, PA10, PA13	5
Chat log analysis	PA2, PA5, PA6, PA13, PA26	5
User behaviour analysis	PA15, PA24	2
Usability analysis	PA22	1
User satisfaction assessment form	PA5	1

Table 8 presents the papers categorized by evaluation method. The paper ID and amount of papers are shown.

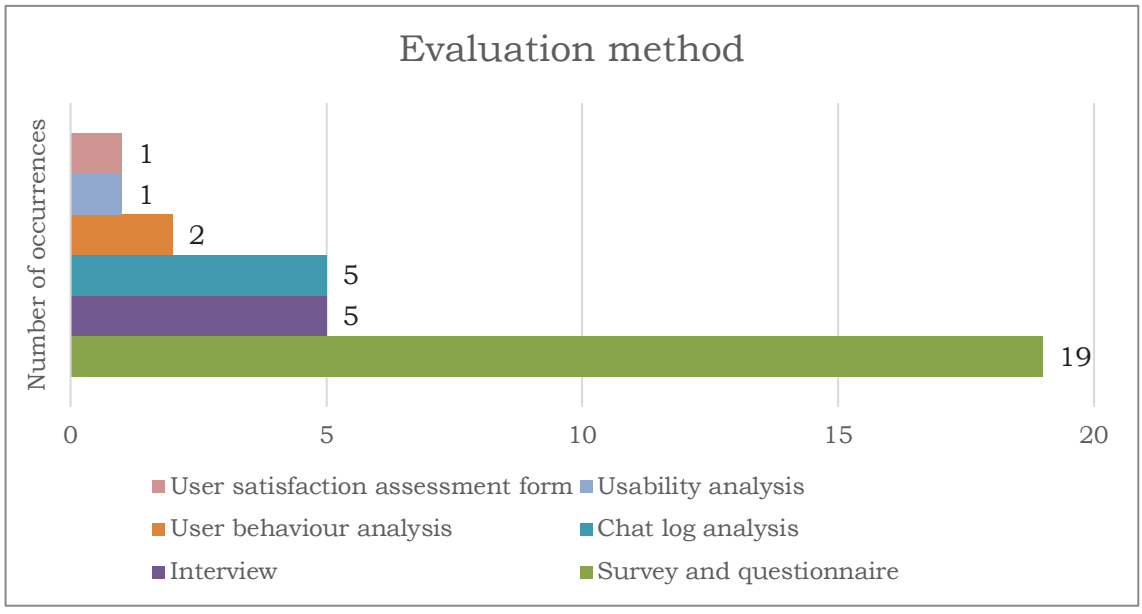

Figure 7. Frequency of research published according to evaluation method.

Figures 7 displays number of occurrences per evaluation method in graph manner. From Table8 and Figure 7, both point out that most of them used survey and questionnaire as an evaluation instrument as it was shown that 19

papers applied it. There were several types of questions using in the survey and questionnaires. Sometimes, it could be in scale manner asking participants to provide the rate with the specific range. In the other hand, it could be open questions which let users input freely their minds. User satisfaction assessment, used by PA5, was also another type of survey but in more formative and well-structured way.

Interview was the second popular method. Some papers acquired both survey and summary. It could be done in several perspective as well. PA7 interviewed the developers and manager of the product. PA4 interviewed the expert and asked for their opinions. Whereas, the remaining three (PA1, PA10, PA13) interviewed users or experiment's participants.

Chat log was also the second popular method and used for data analysis. Various studies chose this method to measure the accuracy of the information provided by the chatbot as well as quality of the conversation. Moreover, 2 papers acquired user behavior analysis. PA24 assessed how long the users spent time with the system. Usability analysis was applied by PA22 in the same manner as user experience analysis.

5.3.2 Evaluation Perspective

In this chapter, what had been measured in the primary studies are presented. Similar to evaluation method, it was possible that one paper could measure several things and presented them. The table 6 represents the different perspective of evaluation as well as how often it appeared in literature.

Table 9. Evaluation perspective and papers categorization.

Evaluation perspective	Paper ID	Number of occurrences
Amount of time with system	PA1, PA24	2
Effectiveness	PA1, PA15, PA17	3
User satisfaction	PA1, PA4, PA6, PA8, PA13, PA14, PA17, PA18, PA24, PA26	10
UX design	PA2	1
Degree of realism	PA3, PA14, PA19, PA20, PA21. PA23, PA25	7
Quality	PA5, PA8, PA16, PA18, PA21, PA25	6
User Engagement	PA5, PA10, PA17	3
Accuracy	PA5, PA6, PA12, PA17, PA20	5
Business Value	PA7, PA8, PA11, PA17, PA20, PA23	6
User Experience	PA9, PA10, PA11, PA12, PA16, PA22	6
Usability	PA11, PA21, PA23	3
Social Presence	PA14, PA19, PA20	3

Table 9 presents the papers categorized by year of publication. The paper ID and amount of papers are shown.

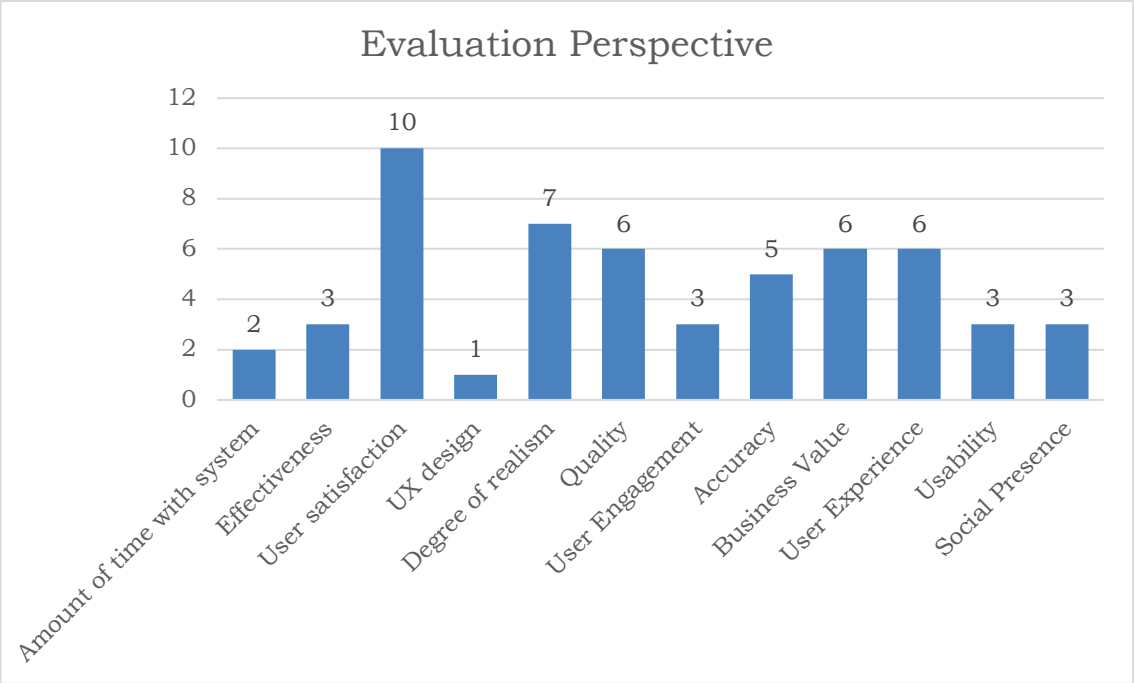

Figure 8. Frequency of research published according to evaluation perspective.

According to table 9 and figure 8, total of 12 evaluation perspectives were presented in primary studies. Many studies mentioned that user satisfaction was evaluated but in a general way. In those, they did not go deeply in details in which perspectives of satisfaction were assessed. Therefore, there were several papers marked with user satisfaction evaluation. The natural of chatbot was also much considered by the researchers as 7 papers out of 26 already adopted it. More human-like meant more positive feedback from users. Most of time, the evaluation of this was done by post-survey.

Quality, business value and user experience were tied with 6 papers mentioned about them. Quality assessment was mostly done by survey and chat log analysis. Business value also gained more attention in research area. It was about the worth of implementing chatbot systems comparing to cost that companies need to pay. It was done by manager interview (PA7) and feedback from users. User experience was also being considered as it could provide fruitful feedback of the system.

Accuracy of the system was about how correct the data from chatbot system to users. Many were assessed by data analysis as well as the survey done by user after the experience with the system. It was stated that 5 papers chose to evaluate the accuracy. Amount of time spending with system, effectiveness, user

engagement, usability and social presence were also evaluated but not in so many.

6. Discussion

The purpose of this study was to discover the research trend and presence of user satisfaction evaluation of chatbot system used in customer service context. Prior studies in the recent years were gathered and categorized in order to find out answers to research questions. The research approach chosen was systematic mapping study. There were 26 articles selected as primary studies. Following the result of study from previous chapter, this chapter discusses the answer the of each sub-question respectively then main research question. Second part provides discussion on research methodology.

6.1 User satisfaction of chatbots system as presented in prior literature

The sub-question Q1 considered the amount of research of each year and hot did it change and it set as:

Q1: How did the amount of research change according to time?

The date range of searching was limited to be only 2010-2020 in order to focus only current research. According to the result from section 4.1, the trend shows that more literature published in recent years. Starting from 2018, the number of papers were dramatically increased.

In the earlier time, 2010 – 2015, there was only one article published in 2013. It can be calculated as 4%. Moreover, there was no studies published in 2014 and 2015. This finding indicated that the chatbot system in terms of user satisfaction was not popular at that time. In 2016, there was another paper published while another one in 2017. Despite that, in 2018, the topic became massively popular. 10 papers or 38% of primary studies were published. The trend still lasted in 2019 which about the same amount (9 papers or 35%) of papers presented. Continue to the year that this study had been done (May, 2020), the trend is still the same as 4 papers or 15% published.

According to prior literature, it is mentioned the chatbots system started receiving massive attention in 2015 due to the recent technology transmission. Artificial Intelligence field is also popular technology and becomes the center of attention among commercial product developers as well as researchers. Furthermore, prior to previous studies, the research trend is still focused on technical perspective while this study focuses on user satisfaction as well as evaluation. Therefore, it supports the finding that more primary studies were conducted after 2018 and the topic was still fresh in academic area. However, the trend showed the potential of increase in the future.

The sub-question Q2 considered the research methodology. It was focused on how the research had been conducted and the motivation of the research.

Q2: How and why did the researchers conduct the research?

When considering the research environment, there were 2 categories presented in this study; experimental and case study. Experimental research was conducted in the controlled environment. Participants were aware that they

were being observed. Usually, they were given easy tasks to complete and might be requested to do surveys and interviews for system evaluations. On the other hand, case study was the observation from the actual usage. The participants were requested to provide information regarding the real situation. It meant they did not know anything about study when they performed the task. Moreover, the data could be from developer's point of view as well as the chat history stored in the product.

The study indicated that most of the research adopted experimental approach with number of 73% of overall primary studies. In the experimental study, researchers could control the research environment in order to reduce the uncontrollable and disinterested factors and fully focused on primary factors presented in the research.

The objectives of primary studies were divided into 3 groups; new system proposed, technology and algorithm improvement and existing system evaluation. There was no outstanding category that caught the eyes of researchers. The number in each group was about the same to each other.

Considering together two perspectives, objective and environment, the results indicated that only for existing system evaluation adopted both case study and experimental. It is possibly reasonable as case study can work only in developed system. Interesting part was that even for the existing system evaluation, experimental study was still preferred in some cases as the tasks and data could be easily framed. Nevertheless, other two objectives still used experimental studies. Technology improvement could be adopted this approach as it was fully focused on interested factors while it was impossible to use case study for new system proposed.

The sub-question Q3 was focused on evaluation part of chatbot system regarding user satisfaction.

> Q3: How did the existing literature evaluate the user satisfaction of chatbot service?

Two objects were studied to answer this research question. Those two were evaluation methods and evaluation perspectives. The combination of these two could give the better picture of the answer to the research question.

For evaluation methods, the research was adopted various techniques including surveys and questionnaires, interview, chat log analysis, user behavior analysis, usability analysis and user satisfaction assessment form. Even assessment form was one kind of survey, it was placed in different category.

The most commonly used method was surveys and questionnaires which 19 out of 26 primary studies adopted it. It was calculated as 73%. Regarding the prior literature, even in the similar topic, evaluation but in different perspective, surveys is also the most commonly used one. The reason is that the respondents could provide the direct response regarding the research questions. The second most used method was interview and chat log analysis with 5 papers adopted each. For interviews, it could be done with users in the same manner as surveys and questionnaires or it could be done with other people, for instance, developers or companies in order to measure business value. Chat log analysis

was performed in various ways, for instance, the accuracy of system's response or time length spending to complete the given task. However, only few considered usability measurements.

Second object, evaluation perspective, was categorized during the implementation of this study. In one primary study, there could be multiple aspects evaluated. The most favorable one among researchers was user satisfaction. Even this was research topic, there was no details described in most of the previous studies. Naturalness of the chatbot had been considered a lot in the recent research.

The main research question was set to discover the research trend of chatbot system in user satisfaction's point of view.

> How is the topic of user satisfaction of chatbot system in customer service presented in the prior academic research?

According to the research findings, the topic was still new and fresh in academic research area as it had not been studied for many years. However, from the high attention among researchers and technology transmission, there was an opportunity that chatbots would become more popular in future research. In deeper details, as there are already presented papers from technology perspectives, more studies from user satisfaction perspective were lacking behind. Moreover, there were only limited research of evaluation dealing with the production. Usually, in lab environment, participants were given the easy task which may not be valid in real situation, especially in the customer service context. When it came to evaluation methods, details of user satisfaction were not well described, for example, in which aspect that user satisfaction had been measured.

6.2 Research approach

The systematic mapping study was chosen in this study as it could be the initial point motivating the future research as well as being the guidance for researchers. It helped summarize the academic papers in order to identify the gap of research area. The study strictly followed guideline mentioned in chapter 3.2.

Every step was performed by individual researcher despite the recommendation to have several researchers to help in evaluation. According to prior literature, the research question shall be set towards the discovery of research trend. Instance scopes are venue, year of publication, journals, methodology and etc. This study did not include every suggested scheme but only year of publication and methodology. Moreover, there was additional scheme, evaluation methods, included in the study. It was the specific scheme towards research topic. With these three themes, it was already enough to answer the main research question.

For conducting the search step, the database search strategy was chosen and Google Scholar was utilized. Google Scholar is also linked to broad scientific databases. This led to timesaving searching process. Moreover, the website provided sorting feature which ranked the result based on relevance. The ranking algorithm considered weighing full text, where it was published authors,

how often it has been cited (Google, n.d.). The access grant was provided by university of Oulu. The search string consisted of four parts; 'chatbots', 'satisfaction', 'customer service' 'case study/experimental study'. According to prior literature, 'conversational agent' is interchangeable term of 'chatbots'. Therefore, total of four search strings were used. There was a great amount of results in return.

Pre-set criteria using in this study may easily introduced the bias. For instance, only English papers were selected. The papers conducted in foreign language might be excluded. Therefore, the mapping might not be presented well. For next step, data extraction and analysis were performed by individual person. All the categories and information were based on personal experience and understanding. Therefore, bias could also be introduced here.

Regarding prior studies, there are several of common threats to validity considering the research method of systematic mapping study. The examples include researcher bias, articles selection criteria and etc. In this study, there were number of threats to be considered. The details of them were listed below.

1. **Researcher bias**. As mentioned earlier, in this research methodology, the bias could be easily existed. Starting from papers section to data extraction and analysis. The result was heavily based on one's opinion, experience and understanding. It is suggested by having more than one researcher working on research in order to avoid this threat.
2. **Papers Accessibility.** The access grant was provided by University of Oulu. However, there were some articles which could not accessed at the time this study had been conducted.
3. **Pre-set criteria.** The pre-set criteria might discourage some of primary studies. Studies in other language than English were not included. Therefore, the study could not be able to represent the global data as some papers might conduct in foreign language and be valid in specific location. Moreover, the decision of selecting only first 50 relevant papers from each search string could also block the valid primary studies to be chosen.

In conclusion, the objectives of systematic mapping studies research methodology were to identify the state of art of current research, draw the picture of research trend, identify the gap, structure the research area. In this study, the methodology was adopted and it could help answer the research question and achieved the research's goal.

7. Conclusion

This study was about user satisfaction of chatbot service in customer service context. The research method was called systematic mapping study. This methodology helps scope the structure of research area in specific topic as well as provide the guidance for future research. This method has been paid high attention by the researchers in recent years (Peterson et al., 2015). The study was conducted on January 2020 – June 2020.

Research method of systematic mapping study (SMS) was selected over systematic review (SR). Systematic mapping study provides broader result which using more primary studies but not in deeper details while systematic review looks insight into details of evidence. Therefore, systematic mapping study provides big picture of research area and is capable with research questions.

In order to get insight into the research area, the articles that published in recent years (2010-2020) were assessed and Google Scholar was utilized as a primary data source. From the search result, 41360 articles were found. The data selection was performed with pre-defined criteria. At the end, 26 articles were selected as primary studies. After that, all articles were evaluated and categorized in various theme based on research questions. At the end, the analysis and discussion were performed.

The main research question was how the topic of user satisfaction of chatbot service in customer service is presented in the prior academic research. In order to answer this research question, it was distributed into 3 sub-questions. Q1 How did the amount of research change according to time? Q2 How and why did the researchers conduct the research? Q3 How did the existing literature evaluate the user satisfaction of chatbot service? All primary studies were categorized from each sub-question perspective then all analysis data were gathered together to answer main research question.

The contribution of this study was to draw of the picture of current status of research trend in order to encourage the researchers conduct the future research. The result from study indicated that the topic is still fresh in academic area while some perspectives are still lacking of attention. The research regarding the evaluation from user's perspective is encouraged. The evaluation of user satisfaction details may not be enough in current state. Moreover, especially in customer service, the case study research which conducted with production application is still lacking and is highly recommended for future research.

There are number of limitations of this study. First threat was the writer's bias. This thesis had been done by individual researcher. Starting from study selection step to categorizing, all findings were heavily based on individual experiences and opinion. Another apparent limitation was accessibility to prior studies. The grant to access the database was provided by University of Oulu. There could be some literature which could not be accessed by the university's grant. Moreover, pre-set criteria could be also threat to validity. For example, the language of prior studies was limited to be only English. Lastly, the quality of primary research could not be guaranteed.

As per the findings of this study. More future research regarding the user satisfaction evaluation is highly recommended. The details such as satisfaction metrics and more reliable methods are possibly absent. Regarding the customer service theme, the real case study should be more presented than in lab environment.

Furthermore, the future research suggested is the improvement of this study. Other scheme shall be performed as there are not so many research conducted in this topic with systematic mapping study (SMS) approach. Another suggestion is systematic review (SR) research as systematic mapping study is commonly starting point for SR.

The outcomes as well as research procedure using in this study are hoped to be beneficial for other research.

8. Bibliografía

Adam, M., Wessel, M., & Benlian, A. (2020). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 1-19.

Akhtar, M., Neidhardt, J., & Werthner, H. (2019, July). The Potential of Chatbots: Analysis of Chatbot Conversations. In *2019 IEEE 21st Conference on Business Informatics (CBI)* (Vol. 1, pp. 397-404). IEEE.

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, 85, 183-189.

Baier, D., Rese, A., & Röglinger, M. (2018, December). Conversational User Interfaces for Online Shops? A Categorization of Use Cases. In *ICIS*.

Bello, M. J. G. (2019). Cloud-based Conversational Agents for User Acquisition and Engagement.

Catapang, J. K., Solano, G. A., & Oco, N. (2020, February). A Bilingual Chatbot Using Support Vector Classifier On an Automatic Corpus Engine Dataset. In 2020 International Conference on *Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 187-192). IEEE.

Chung, M., Ko, E., Joung, H., & Kim, S. J. (2018). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*.

Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., & Zhou, M. (2017, July). Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL 2017, System Demonstrations* (pp. 97-102).

Dahiya, M. (2017). A tool of conversation: Chatbot. *International Journal of Computer Sciences and Engineering*, 5(5), 158-161.

Dibitonto, M., Leszczynska, K., Tazzi, F., & Medaglia, C. M. (2018, July). Chatbot in a campus environment: design of LiSA, a virtual assistant to help students in their university life. In *International Conference on Human-Computer Interaction* (pp. 103-116). Springer, Cham.

Diederich, S., Brendel, A. B., Lichtenberg, S., & Kolbe, L. (2019). DESIGN FOR FAST REQUEST FULFILLMENT OR NATURAL INTERACTION? INSIGHTS FROM AN EXPERIMENT WITH A CONVERSATIONAL AGENT.

Denyer, D., & Tranfield, D. (2009). Producing a systematic review.

Elmasri, D., & Maeder, A. (2016, October). A conversational agent for an online mental health intervention. *In International Conference on Brain Informatics* (pp. 243-251). Springer, Cham.

Elsholz, E., Chamberlain, J., & Kruschwitz, U. (2019, March). Exploring Language Style in Chatbots to Increase Perceived Product Value and User Engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (pp. 301-305).

Følstad, A., & Skjuve, M. (2019, August). Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (pp. 1-9).

Ghose, S., & Barua, J. J. (2013, May). Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor. In *2013 international conference on informatics, electronics and vision (ICIEV)* (pp. 1-5). IEEE.

Gnewuch, U., Morana, S., Adam, M., & Maedche, A. (2018). Faster is not always better: understanding the effect of dynamic response delays in human-chatbot interaction.

Google. (n.d.). Google Scholar About, Retrieved May 23, 2020 from https://scholar.google.com/intl/en/scholar/about.html

Heo, M., & Lee, K. J. (2018). Chatbot as a new business communication tool: The case of naver talktalk. *Business Communication Research and Practice*, 1(1), 41-45.

Hung, V., Elvir, M., Gonzalez, A., & DeMara, R. (2009, October). Towards a method for evaluating naturalness in conversational dialog systems. *In 2009 IEEE International Conference on Systems, Man and Cybernetics* (pp. 1236-1241). IEEE.

Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., ... & Akkiraju, R. (2018, April). 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).

Hwang, S., Kim, B., & Lee, K. (2019, July). A data-driven design framework for customer service chatbot. In *International Conference on Human-Computer* Interaction (pp. 222-236). Springer, Cham.

Io, H. ., & Lee, C. B. (2017, December). Chatbots and conversational agents: A bibliometric analysis. *In 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 215-219). IEEE.

Johari, N. M., Zaman, H. B., & Nohuddin, P. N. (2019, November). Ascertain Quality Attributes for Design and Development of New Improved Chatbots to Assess Customer Satisfaction Index (CSI): A Preliminary Study. In *International Visual Informatics Conference* (pp. 135-146). Springer, Cham.

Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. (2003). Five steps to conducting a systematic review. *Journal of the royal society of medicine*, 96(3), 118-121.

Koetter, F., Blohm, M., Kochanowski, M., Goetzer, J., Graziotin, D., & Wagner, S. (2018). Motivations, classification and model trial of conversational agents for insurance companies. *arXiv preprint arXiv:1812.07339*.

Kopp, S., & Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds*, 15(1), 39-52.

Kuligowska, K. (2015). Commercial chatbot: performance evaluation, usability metrics and quality standards of embodied conversational agents. *Professionals Center for Business Research*, 2.

Muischnek, K., & Müürisep, K. (2018, September). Collection of Resources and Evaluation of Customer Support Chatbot. In *Human Language Technologies–The Baltic Perspective: Proceedings of the Eighth International Conference Baltic HLT 2018* (Vol. 307, p. 30). IOS Press.

Nguyen, X., Tran, H., Phan, H., & Phan, T. (2020). Factors influencing customer satisfaction: The case of Facebook Chabot Vietnam. *International Journal of Data and Network Science*, 4(2), 167-178.

Peras, D. (2018). Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, 89-97

Perski, O., Crane, D., Beard, E., & Brown, J. (2019). Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. *Digital health*, 5, 2055207619880676.

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008, June). Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12* (pp. 1-10).

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64, 1-18.

Ren, R., Castro, J. W., Acuña, S. T., & de Lara, J. (2019). Usability of chatbots: A systematic mapping study. *In Proc. 31st Int. Conf. Software Engineering and Knowledge Engineering* (pp. 479-484).

Romero-Charneco, M., Casado-Molina, A. M., & Alarcón-Urbistondo, P. (2018). Channels of social influence for decision making in restaurants: A case study. *Dos Algarves: A Multidisciplinary e-Journal*, 32, 54-76.

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.

Shawar, B. A., & Atwell, E. (2007, April). Different measurement metrics to evaluate a chatbot system. In *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies* (pp. 89-96).

Shawar, B. A., & Atwell, E. (2007, January). Chatbots: are they really useful?. *In Ldv forum* (Vol. 22, No. 1, pp. 29-49).

Thomas, N. T. (2016). An e-business chatbot using AIML and LSA. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2740-2742). IEEE.

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.

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).

Appendix A. List of primary studies

Paper ID	Paper name	Author name(s)	Year of publication
PA1	A conversational agent for an online mental health intervention.	Elmasri, D., & Maeder, A.	2016
PA2	A data-driven design framework for customer service chatbot	Hwang, S., Kim, B., & Lee, K.	2019
PA3	AI-based chatbots in customer service and their effects on user compliance	Adam, M., Wessel, M., & Benlian, A.	2020
PA4	Ascertain Quality Attributes for Design and Development of New Improved Chatbots to Assess Customer Satisfaction Index (CSI): A Preliminary Study.	Johari, N. M., Zaman, H. B., & Nohuddin, P. N.	2019
PA5	Bilingual Chatbot Using Support Vector Classifier On an Automatic Corpus Engine Dataset	Catapang, J. K., Solano, G. A., & Oco, N.	2020
PA6	Channels of social influence for decision making in restaurants: A case study	Romero-Charneco, M., Casado-Molina, A. M., & Alarcón-Urbistondo, P.	2018
PA7	Chatbot as a new business communication tool: The case of naver talktalk	Heo, M., & Lee, K. J.	2018
PA8	Chatbot e-service and customer satisfaction regarding luxury brands	Chung, M., Ko, E., Joung, H., & Kim, S. J.	2018
PA9	Chatbot in a campus environment: design of LiSA, a virtual assistant to help students in their university life	Dibitonto, M., Leszczynska, K., Tazzi, F., & Medaglia, C. M.	2018
PA10	Chatbots for customer service: user experience and motivation	Følstad, A., & Skjuve, M.	2019
PA11	Cloud-based Conversational Agents for User Acquisition and Engagement.	Bello, M. J. G.	2019
PA12	Collection of Resources and Evaluation of Customer Support Chatbot	Muischnek, K., & Müürisep, K.	2018

Paper ID	Paper name	Author name(s)	Year of publication
PA13	Conversational User Interfaces for Online Shops? A Categorization of Use Cases	Baier, D., Rese, A., & Röglinger, M.	2018
PA14	DESIGN FOR FAST REQUEST FULFILLMENT OR NATURAL INTERACTION? INSIGHTS FROM AN EXPERIMENT WITH A CONVERSATIONAL AGENT.	Diederich, S., Brendel, A. B., Lichtenberg, S., & Kolbe, L.	2019
PA15	Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study	Perski, O., Crane, D., Beard, E., & Brown, J.	2019
PA16	Examining the customer experience of using banking Chatbots and its impact on brand love: the moderating role of perceived risk. Journal of internet Commerce	Trivedi, J.	2019
PA17	Exploring Language Style in Chatbots to Increase Perceived Product Value and User Engagement	Elsholz, E., Chamberlain, J., & Kruschwitz, U.	2019
PA18	Factors influencing customer satisfaction: The case of Facebook Chabot Vietnam	Nguyen, X., Tran, H., Phan, H., & Phan, T.	2020
PA19	Faster is not always better: understanding the effect of dynamic response delays in human-chatbot interaction	Gnewuch, U., Morana, S., Adam, M., & Maedche, A.	2018
PA20	Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions	Araujo, T	2018
PA21	Motivations, classification and model trial of conversational agents for insurance companies.	Koetter, F., Blohm, M., Kochanowski, M., Goetzer, J., Graziotin, D., & Wagner, S.	2018
PA22	Superagent: A customer service chatbot for e-commerce websites	Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., & Zhou, M	2017
PA23	The analysis of customer satisfaction factors which influence chatbot acceptance in Indonesia	Sanny, L., Susastra, A., Roberts, C., & Yusramdaleni, R.	2020

Paper ID	Paper name	Author name(s)	Year of publication
PA24	The Potential of Chatbots: Analysis of Chatbot Conversations	Akhtar, M., Neidhardt, J., & Werthner, H.	2019
PA25	Touch your heart: a tone-aware chatbot for customer care on social media	Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., ... & Akkiraju, R.	2018
PA26	Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor.	Ghose, S., & Barua, J. J.	2013