EPiC Series in Computing
Volume 84, 2022, Pages 100–116
Proceedings of the Society 5.0 Conference 2022
- Integrating Digital World and Real World
to Resolve Challenges in Business and Society

A Computational Literature Analysis of Conversational AI Research with a Focus on the Coaching Domain

Charuta Pande[1,2][0000-0001-6530-5401], Hans-Georg Fill[2][0000-0001-5076-5341] and Knut Hinkelmann[1][0000-0002-1746-6945]

[1] FHNW University of Applied Sciences and Arts Northwestern Switzerland,
Intelligent Information Systems Research Group, Riggenbachstrasse 16, 4600, Olten, Switzerland
charuta.pande@fhnw.ch, knut.hinkelmann@fhnw.ch

[2] University of Fribourg, Research Group Digitalization and Information Systems,
Bd de Pérolles 90, 1700 Fribourg, Switzerland
charuta.pande@unifr.ch, hans-georg.fill@unifr.ch

Abstract

We conduct a computational analysis of the literature on Conversational AI. We identify the trend based on all publications until the year 2020. We then concentrate on the publications for the last five years between 2016 and 2020 to find out the top ten venues and top three journals where research on Conversational AI has been published. Further, using the Latent Dirichlet Allocation (LDA) topic modeling technique, we discover nine important topics discussed in Conversational AI literature and specifically two topics related to the area of coaching. Finally, we detect the key authors who have contributed significantly to Conversational AI research and area(s) related to coaching. We determine the key authors' areas of expertise and how the knowledge is distributed across different regions. Our findings show an increasing trend and thus, an interest in Conversational AI research, predominantly from the authors in Europe.

Keywords: Conversational AI, Chatbots, Conversational Agents, Coaching, Computational Literature Analysis, Latent Dirichlet Allocation

1 Introduction

Technology facilitated training, learning and coaching have been identified as an important area of Society 5.0. In Society 5.0, technology acts as an effective enabler for affordable and accessible

education[*], and personalized coaching and training have the potential to improve work and living standards[†]. In areas where coaches are not accessible, the digitalization of coaching makes a significant contribution to society. Coaching is an interactive activity. Standard interactions like customer service (Følstad & Skjuve, 2019) and frequently asked questions (FAQ) (De Lacerda & Aguiar, 2019) have been effectively digitalized using conversational agents. We would like to investigate if the delegation of complex and non-standard interactions like coaching can also be achieved using conversational agents.

Conversational AI is the field of research that encompasses studies around software agents, also known as conversational agents, that can carry out conversations in natural language with humans (Jurafsky & Martin, 2020). Literature distinguishes conversational agents into two categories based on the goal of the conversation – those that carry out a social conversation on various topics and those that have a specific goal or a task to perform (Gao et al., 2019; Hussain et al., 2019). The term "chatbot" usually refers to the former category. Over time, the terms "chatbot" and "conversational agent" have been interchangeably and generally used to represent a software agent that can conduct a conversation in natural language (Jurafsky & Martin, 2020).

Conversational agents are applied in several domains like healthcare, finance, education, tourism, and many more. Conversational AI includes studies in specialized fields like design techniques, anthropomorphic aspects (Pamungkas, 2019; Rapp et al., 2021), user experience (Skjuve & Brandzaeg, 2019), evaluation methods (Maroengsit et al., 2019), and neural approaches (Gao et al., 2019), to name a few. We would like to explore coaching as an application area of Conversational AI.

A preliminary search showed that coaching is applied in domains like healthcare (Wolever et al., 2017), education (Fletcher & Mullen, 2012), organizational training (Grant et al., 2010) and sports (Gale, 2017). Since it is not evident how extensively conversational agents have been used in coaching, we want to understand the state-of-the-art in Conversational AI research. We want to identify the main venues of publications in Conversational AI, so that we can create a foundation for future research in "Conversational AI in coaching". To strengthen it further, we would also like to identify how the expertise in Conversational AI is distributed across regions and what are the prominent areas of expertise. With these objectives in mind, we identified the following research questions:

RQ1: How has the number of publications in Conversational AI evolved over time?

RQ2: What are the main venues/outlets where research on Conversational AI has been published?

RQ3: What are the major topics, areas of application, and domains considered in Conversational AI research?

RQ4: Which topics are investigated in Conversational AI for coaching?

RQ5: How many authors are active in Conversational AI research and in coaching, and what are their areas of expertise?

The specific contribution of this paper is a *computational literature analysis* of all research fields in Conversational AI using text-mining approach and is performed on *two scientific sources* – DBLP[‡] and Scopus[§]. The current research is exploratory in nature and aims to discover areas where future research related to Conversational AI in coaching can be focused. Further, this work results in the identification of top venues that can be interesting for publishing future research, the detection of major topics using a topic modeling technique and discovering the key authors in the field of Conversational AI. We first focus on a quantitative analysis of Conversational AI research in general, followed by a content and author analysis of Conversational AI research, and then more specifically, content and author analysis of the research related to Conversational AI in coaching.

[*] https://www.up.ac.za/faculty-of-engineering-built-environment-it/news/post_2953210-society-5.0-offers-the-possibility-of-affordable-accessible-education-for-all-up-expert

[†] https://www.strategyand.pwc.com/m1/en/ideation-center/ic-research/2021/assets/study/Society50-Downloadable.pdf

[‡] https://dblp.org/

[§] https://www.scopus.com/

The rest of the paper is organized as follows: in Section 2 we describe related work on literature reviews on conversational agents and the followed methodology in Section 3. Section 4 describes the quantitative analysis for Conversational AI in general and the content analysis is described in Section 5. The author analysis is described in Section 6, followed by a discussion of the findings in Section 7, and finally, the conclusion and future work.

2 Related Work

Systematic Literature Review (SLR) is an effective approach to explore, understand, and organize literature around precise research questions and derive insights from scientific literature. However, performing SLR manually requires significant time, effort and possibly collaboration (Okoli, 2015). Computational approaches for literature review allow analysis of much larger amount of data as well as reduce the time and effort involved (Asmussen & Møller, 2019).

Previous literature reviews on Conversational AI include SLRs confined to specific domains, e.g., chatbots in business (Bavaresco et al., 2020), text-based chatbots (Rapp et al., 2021), chatbots in education (Pérez et al., 2020) and chatbots in tourism (Calvaresi et al., 2021). Other analyses include Bibliometric Analysis of chatbots and conversational agents (Bernardini et al., 2018; Io & Lee, 2017) that use clustering and term co-occurrence techniques on data from a single database – either Scopus or Web of Science. An SLR by Caldarini et al. (2022) focuses on state-of-the-art of chatbots and deep learning algorithms. In this work, we abstract from the various domains and explore Conversational AI as a research area to gather deeper insights into this field.

3 Methodology

We followed a computational approach for SLR called Computational Literature Review (CLR), based on a six step roadmap proposed by Antons et al. (2021). CLR uses computational algorithms to analyze literature as opposed to the traditional SLR, where analysis is mostly performed manually. Figure 1 outlines the Data Preparation and the CLR process applied in this research. We describe the process in the following sections[**].

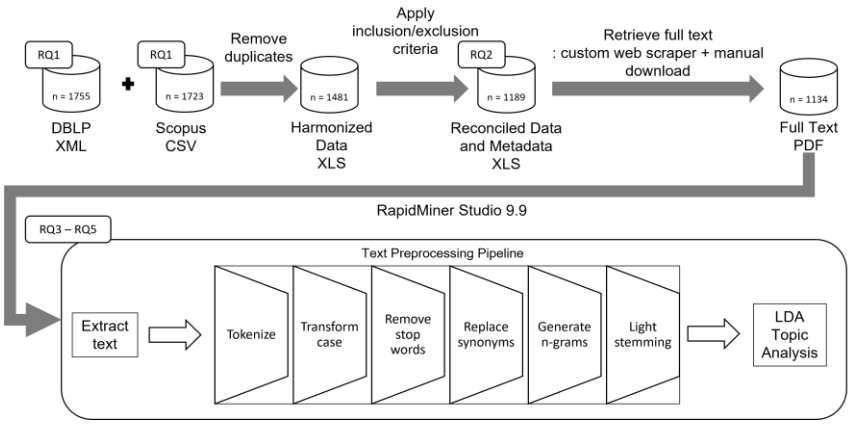

Figure 1: Data Preparation and CLR pipeline based on data from DBLP and Scopus

[**] The details of the data analyzed in this research including the document titles, the corresponding DOIs, authors and the RapidMiner process to extract topics is available at https://doi.org/10.5281/zenodo.5913880

3.1 Data Collection and Preparation

The main sources for data collection were DBLP, a specialized freely accessible database for computer science publications and Scopus, a multi-disciplinary database that includes publications from different fields of science. Both databases are known to index a high number of unique publications (Cavacini, 2015) and it is reasonable to consider them together as specialized research is indexed only in DBLP and not in Scopus. A keyword search using the terms “chatbot” and “conversational agent” was performed in early 2021 in DBLP to identify research papers until the end of the year 2020. DBLP allows search only on metadata, hence the search was performed as a combined search including authors, venues, and publication titles. This search returned 1755 papers. To broaden the scope of the study, a search with the same keywords and until the end of 2020 was performed in Scopus, which resulted in 1723 papers. The search query in Scopus was (TITLE ("conversational agent") OR TITLE (chatbot)) AND PUBYEAR < 2021. The search keywords “chatbot” and “conversational agent” were chosen as we observed that they are the most generic terms used to address conversational agents in the literature.

The analysis for RQ1 was performed separately on the individual databases. For further analysis, the data from the two databases were harmonized. The overlap between DBLP and Scopus data was resolved by filtering out the duplicate data. For our analysis, we considered only research papers submitted to conferences, symposia, journals, and workshops except for preprints from CoRR[††] that can be potentially submitted to a conference. Duplicate data, e.g., preprints that were later submitted to conferences were filtered out. Further, we considered only papers written in English as all major journals have publications in English. Moreover, the overall volume of non-English papers retrieved in our search results was quite low (N=33). Thus, all non-research documents and papers written in languages other than English were filtered out. This resulted in 1189 papers which constituted as data for the quantitative analysis for RQ2. The metadata was retrieved by parsing the results from DBLP (XML) and Scopus (CSV) and was harmonized. Further, missing, or incorrect metadata was manually reconciled. The quantitative analysis for RQ1, RQ2, RQ4, and RQ5 was performed using Pivot Tables in Microsoft Excel. For RQ4 and RQ5, the analysis was further augmented by manually retrieving affiliations and research interests.

A total of 55 papers were not accessible and were excluded from the dataset. After applying the inclusion and exclusion criteria shown in Table 1, 1134 papers constituted the corpora for the analysis of RQ3-RQ5. Next, full-text papers were downloaded in PDF format using a custom web scraper. The full-text analysis was performed using the Text Processing extension of the tool RapidMiner Studio 9.9[‡‡]. This corresponds to a workflow that has been successfully used previously in computational literature analyses (Härer & Fill, 2020; Muff et al., 2022). The aim was to apply topic modeling on the extracted text. Topic modeling is a technique applied as an unsupervised text mining approach to large corpora of unstructured text to detect the underlying topics. Topic modeling is the main task in operationalizing CLR (Antons et al., 2021).

Inclusion criteria	Papers submitted to conferences, journals, and workshops Preprints from arxiv.org Papers between 2016 and 2020
Exclusion criteria	Papers written in languages other than English Duplicate papers Presentations, talks, interviews, keynotes, textbooks Papers that could not be accessed

Table 1: Inclusion and Exclusion criteria for the analysis

[††] https://arxiv.org/
[‡‡] https://rapidminer.com/products/studio/

3.2 Operationalize CLR

The full text of the pdf documents was first extracted by defining regular expressions to include only the content of the paper and exclude details like title page, author/affiliation details, and the reference list. Exceptions in text extraction were handled manually. The extracted text underwent preprocessing steps like tokenization, case transformation, stop word removal, synonym replacement, and bigram generation. We filtered out the terms "chatbot" and "conversational agent" to avoid generating a topic biased towards the search keywords. Terms that are closer and similar to coaching were replaced by "coaching". For example, counseling, tutoring, mentoring, and consultation are each distinctive in nature, yet, can be considered quite similar to coaching in terms of a. the conversations between the chatbot (as a coach/tutor/mentor, etc.) and the user, and b. the goal of the conversation, which is to bring about a change in the state of the mind of the user (Irby et al., 2018).

The Latent Dirichlet Allocation (LDA) method (Blei et al., 2003) was then applied to the preprocessed text to identify the underlying topics in the pdf documents. For this, we applied the LDA implementation in RapidMiner, which uses the MALLET[§§] toolkit. The basic assumption of LDA is that every document in a document collection corresponds to several abstract topics with a certain probability. The actual topics are hidden and are interpreted from the word collection belonging to the topic. Thus, each topic has a specific weight associated with each document. Further, each topic is a collection of top words that frequently occur, the frequency of a word in a topic is denoted as its weight (Blei et al., 2003). The number of topics and the number of words in a topic are configurable. The algorithm uses Gibbs sampling, which allows for configuration of additional hyperparameters that optimize and improve the topic model iteratively (Darling, 2011).

According to Schofield & Mimno (Schofield & Mimno, 2016), topic modeling performs better when light stemming is applied to top words *after* modeling. For light stemming, we applied a custom S-removal stemmer. For stop word removal, synonym replacement and the custom S-stemmer we adopted a "lazy" approach, where the terms to be replaced/removed were not decided before hand but identified iteratively by observing the output of the algorithm. After each iteration, we analyzed the topic results and finally chose a result that provided a scope for maximum interpretability. Thus, the intention of this exercise was to balance a tradeoff between information loss and interpretability and avoid over-tuning of the topics by replacing all possible synonyms or plural forms.

We, thus, identified the top nine clearly distinguishable topics with top twenty words in each topic. The topic interpretations are described in sections 5.1 and 5.2. The interpretation of the topics was performed by identifying the most concise label (a collection of terms) that fits most words in that topic. Additionally, the papers associated with a topic with high confidence were examined to precisely interpret the context and the terms in the label. The evaluation is described in section 5.3.

4 Quantitative Analysis of Conversational AI Literature

The first part of our analysis focuses on a general analysis of Conversational AI research and answers RQ1 and RQ2. This includes trend analysis of the entire data and quantitative analysis of data from 2016 to 2020.

4.1 Publication Trend

Since we accessed data from two sources, we decided to perform the trend analysis on individual datasets i.e., before harmonizing the data. This gave us an opportunity to compare and validate the trend against two different scientific sources.

[§§] MAchine Learning for LanguagE Toolkit - http://mallet.cs.umass.edu/topics.php

As can be observed from Figure 2, the trend curves from both databases go almost hand in hand. Based on available data from both datasets, the terms "chatbot" and/or "conversational agent" started appearing in Conversational AI research around the mid-1990s. There is a steep increase in the number of publications after 2015. The number of publications increased by 56.6% in DBLP and 84% in Scopus from 2015 to 2016. Scopus data shows that the publications increased more than ten-fold between 2016 and 2020, indicating a steeply growing interest in the research in Conversational AI. The dip in the DBLP trend curve for the year 2020 reflects the number of publications available in the DBLP database at the time of the search and thus, does not suggest a drop in research interest.

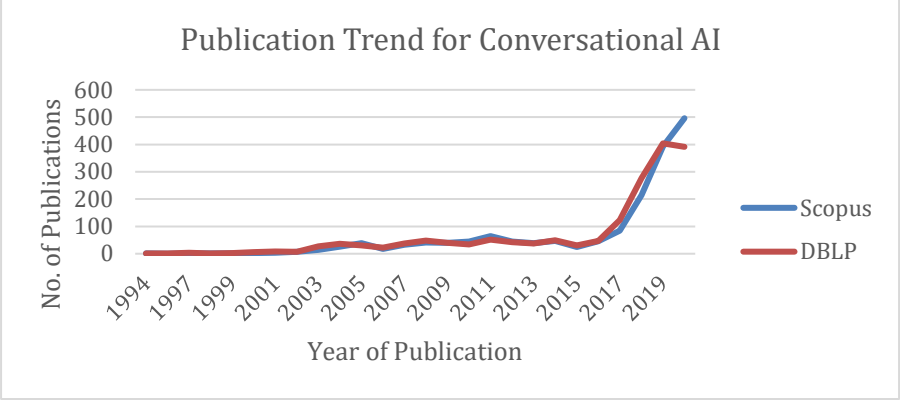

Figure 2: Trend analysis of Conversational AI publications from DBLP and Scopus

Owing to a sharp increase in the number of publications from 2016 onwards, the rest of the analysis for RQ2-RQ5 was performed on the data for a period of five years from the year 2016 to 2020.

4.2 Popular Venues for Conversational AI Research

To identify the popular venues where Conversational AI research has been published, we performed data aggregation on the number of publications from 2016 to 2020. The results have been divided into two categories – A. publications in conferences, workshops, symposia, e-print archives and B. publications in journals. We consider journal publications as a separate category as they differ from conference/workshop/symposia publications in terms of the time taken to review and publish, the possibility to present and discuss the results, and restrictions regarding the number of pages. We included e-print archives in category A., as we observed that in the last five years more publications from archives were submitted to conference/workshop/symposia than to journals. Thus, the analysis was performed on a total of 1040 conference papers and 149 journal publications.

A. Publications in Conferences, Workshops, Symposia, e-Print Archives

Figure 3 shows the top-10 venues of category A and the number of publications for every year. The top venues are – the Computing Research Repository (CoRR), Conference on Human Factors in Computing Systems (CHI), Intelligent Virtual Agents (IVA), Human-Computer Interaction (HCI), CONVERSATIONS, European Conference on Information Systems (ECIS), Intelligent User Interfaces (IUI), Association for the Advancement of Artificial Intelligence (AAAI), Conversational User Interfaces (CUI) and Human-Agent Interaction (HAI). If a conference organizes workshops or symposia, then the corresponding papers are counted together under the conference name or the conference organizer's name.

The maximum publications in Conversational AI for the last five years have been in CoRR (N=80). CoRR is an open access archive for publishing curated e-prints of the latest research in

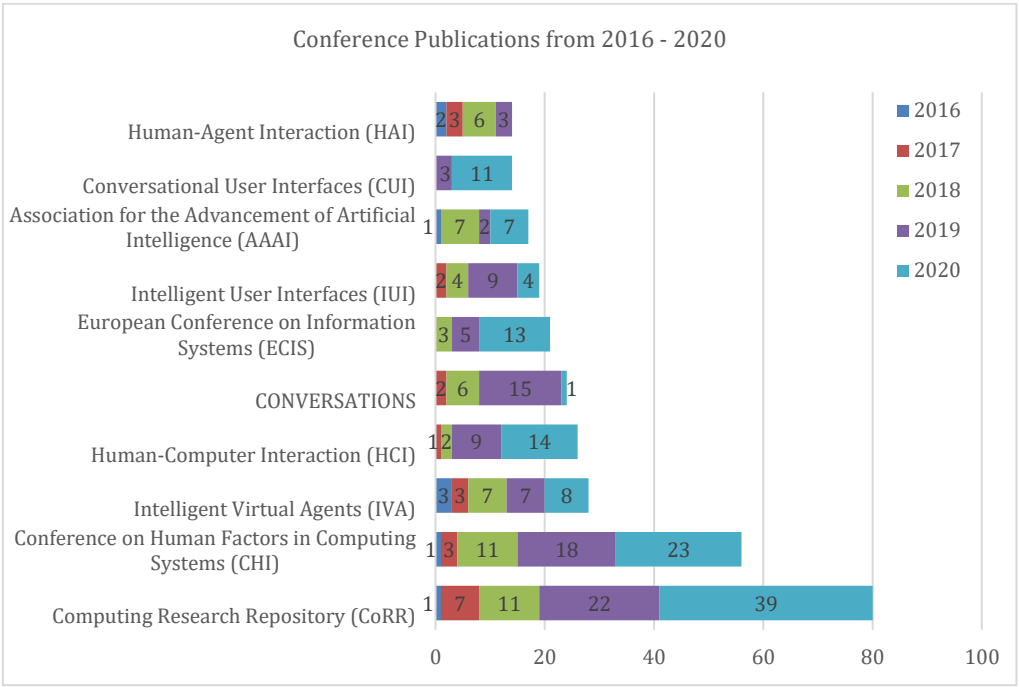

Figure 3: Distribution of conference publications from 2016 – 2020

various areas like Physics, Mathematics, Computer Science, and more. Sutton & Gong (2017) highlight a growing trend of pre-publishing research work as "preprints" in the field of Computer Science, which is also observed in the number of preprints submitted for Conversational AI in CoRR. The results for CoRR in Figure 3 include only preprints that were not yet published in any conference or a journal.

CHI is the most popular conference with a total of 56 submissions across five years - the numbers include also extended abstracts. CHI is followed by the IVA conference (N=28) that includes work on virtual and embodied conversational agents, and the HCI conference (N=26). CONVERSATIONS (N=24) is a multidisciplinary workshop that started in the year 2017 and focuses specifically on chatbot research and design. The publication count for the year 2020 (N=1) is based on the results obtained at the time of the database search. The proceedings for CONVERSATIONS 2020 show that the actual number of publications in that year was 17, which makes the total number of publications over the last five years 40.

In the ECIS conference, the publication count was more than double in 2020 (N=13) compared to 2019 (N=5). IUI (N=19) and AAAI (N=17) address broader research fields of User Interface and Artificial Intelligence and have received a fluctuating number of publications over the last five years. The CUI conference (N=14) was first held in the year 2019 and focuses on speech and text-based conversational user interfaces. The number of publications in CUI more than tripled in 2020 (N=11) from the previous year (N=3). The HAI conference has no publications using the chosen search keywords in 2020, however, a closer look showed that the published papers use different, less common synonyms to the search keywords e.g., "conversational companion" and "human-agent interaction" or simply "HCI".

The data distribution for the conference papers is highly skewed. Although some dedicated venues for Conversational AI research could be identified, they constitute just about 30% of the overall papers submitted to conferences between 2016 and 2020. Almost 70% of the papers were submitted to diverse conferences, emphasizing the inter-disciplinary nature of Conversational AI research. For example, applications for healthcare and life sciences have been published in conferences like EAI Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), IEEE Conference on Life Sciences and Technologies (LifeTech). The theoretical concepts related to Recommender Systems, Natural Language Processing, Information Retrieval have been published in the ACM Conference on Recommender Systems (RecSys), Conference on Empirical Methods in Natural Language Processing (EMNLP), ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR) and the cognitive aspects have been published in Conference on Information and Knowledge Management (CIKM), Conference on Knowledge Engineering and Ontology Development (KEOD). Taking the overall volume of the conference papers into account, the number of publications in 2019 and 2020 was almost twice the number of publications in 2016, 2017, and 2018 combined.

B. Publications in Journals

The popular journals for publishing Conversational AI research can be seen in Figure 4 and are – Computers in Human Behavior, IEEE Access, and the Proceedings of the ACM on Human-Computer Interaction and Engineering. The papers published in the top-3 journals contribute to about 15% (N=22) of the overall papers submitted to all journals, implying that the choice of journal is very specific to the domain, or the discipline of Conversational AI research. A little more than one-third of the journal articles (N=55) are in journals that have only one publication in Conversational AI in the last five years, highlighting that Conversational AI spans several areas of application. The combined number of publications in the years 2019 and 2020 amounts to almost 75% of the publications submitted in journals over the last five years, again emphasizing an increased trend of Conversational AI research.

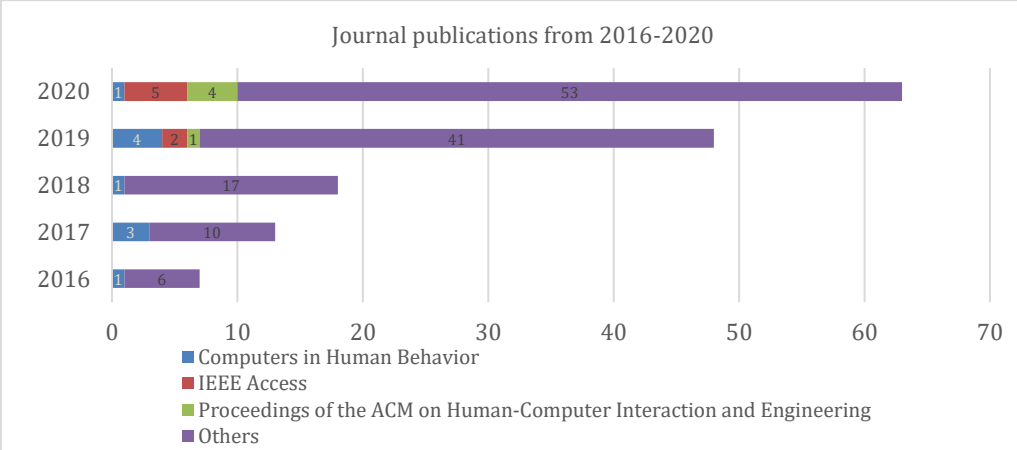

Figure 4: Distribution of journal publications from 2016 - 2020

Overall, it should be noted that the number of publications in 2020 is based on the data available at the time of search and may be more than indicated in Figure 3 and Figure 4 when cross-checked with the individual conferences. To take a closer look at the inter-disciplinary topics in Conversational AI research, the next section elaborates on the full-text analysis of all available publications.

5 Content Analysis

This section describes the results of the LDA algorithm and answers RQ3 and RQ4. Table 2 shows the output of the LDA algorithm - all nine topics, top twenty words in every topic and their corresponding weights. We describe the results and our interpretations by using *italics* style to indicate a term in the top twenty words and underline style to indicate the inferred topic.

Topic 1		**Topic 2**		**Topic 3**		**Topic 4**		**Topic 5**	
word	*wt.*	*word*	*wt.*	*word*	*wt.*	*word*	*wt.*	*word*	*wt.*
intent	2760,0	emotion	2863,0	student	4742,0	customer	883,0	participant	5727,0
model	1741,0	children	1770,0	learning	3809,0	news	442,0	social	3949,0
action	1251,0	participant	1518,0	learner	2369,0	security	398,0	human	3206,0
dialogue	1042,0	eca	1420,0	course	1049,0	repair	363,0	perceived	2739,0
entity	941,0	emotional	1331,0	team	983,0	employees	300,0	design	2268,0
entities	909,0	speech	1204,0	education	860,0	messages	280,0	study	2215,0
user	883,0	interaction	854,0	design	771,0	media	278,0	cas	2179,0
set	722,0	personality	791,0	teacher	735,0	affordances	261,0	interaction	2142,0
dataset	679,0	virtual	759,0	teaching	651,0	government	250,0	response	2133,0
nlu	647,0	behavior	714,0	educational	643,0	slack	246,0	user	2068,0
state	645,0	child	682,0	group	637,0	channels	240,0	research	1881,0
topic	623,0	human	661,0	coaching	586,0	citizens	235,0	cue	1607,0
utterance	620,0	verbal	657,0	pedagogical	578,0	privacy	222,0	trust	1603,0
training	593,0	alexa	561,0	online	455,0	enterprise	204,0	effect	1438,0
crowd	540,0	model	554,0	collaborative	433,0	bots	204,0	self	1341,0
domain	496,0	facial	517,0	cas	409,0	churn	197,0	survey	1313,0
task	479,0	rate	466,0	task	358,0	workers	186,0	service	1305,0
policy	464,0	voice	464,0	study	357,0	call	183,0	presence	1205,0
slot	456,0	affect	428,0	interest	354,0	social	180,0	customer	1054,0
system	442,0	affective	427,0	activities	352,0	organizations	177,0	people	1016,0

Topic 6		**Topic 7**		**Topic 8**		**Topic 9**	
word	*wt.*	*word*	*wt.*	*word*	*wt.*	*word*	*wt.*
question	2952,0	model	10833,0	user	35425,0	health	2806,0
answer	1917,0	response	7288,0	based	12091,0	patient	2096,0
word	1542,0	word	4300,0	information	11121,0	healthcare	955,0
argument	1118,0	training	2816,0	system	10464,0	mental	898,0
query	918,0	context	2814,0	use	10353,0	medical	703,0
aiml	825,0	data	2473,0	data	9597,0	participant	686,0
knowledge	752,0	based	2420,0	question	9220,0	mental_health	595,0
ontology	562,0	dataset	2417,0	using	8793,0	care	551,0
pattern	555,0	dialogue	2076,0	conversation	7373,0	disease	488,0
graph	544,0	matching	1937,0	interaction	7230,0	depression	465,0
similarity	494,0	seq	1892,0	language	7167,0	stress	419,0
system	476,0	set	1891,0	service	6917,0	older	414,0
queries	458,0	sentence	1829,0	result	6818,0	adults	412,0
base	455,0	result	1823,0	work	6735,0	eca	407,0
data	439,0	learning	1775,0	time	6485,0	studies	401,0
sentence	424,0	table	1767,0	human	6378,0	application	399,0
knowledge_base	421,0	human	1761,0	response	5910,0	coaching	390,0
algorithm	420,0	neural	1690,0	systems	5536,0	self	389,0
keyword	413,0	input	1581,0	research	5311,0	symptoms	379,0
question_answer	394,0	trained	1573,0	answer	4995,0	usability	364,0

Table 2: Topics generated by LDA algorithm on papers from 2016 – 2020 ordered by word weight per topic

5.1 Major Topics, Areas of Application and Domains

Topic 1 has *intent*, *model*, *action*, *dialogue*, and *entity* as the most significant words. Together with the rest of the words, this topic was interpreted as training of chatbots for natural language understanding. The bottom-10 words include *nlu*, *training*, and *crowd*, indicating that there is a distinct focus on improving natural language understanding of chatbots through training, and crowd sourcing has been used in several studies for training chatbots.

In Topic 2, *emotion*, *children*, *participant*, *eca*, and *emotional* are dominant. This topic was interpreted as Embodied Conversational Agents (ECA) and focuses on various *behavioral* characteristics like *facial*, *verbal*, *voice*, *personality*, etc., that contribute to the emotions.

With *student*, *learning*, *learner*, *course*, and *team* as the significant words, Topic 3 was clearly interpreted as chatbots in education. Other terms like *group*, *collaborative*, *study*, *online* imply an application of chatbots for online collaborative learning.

Topic 4 was interpreted as usage of chatbots in organizations, where *customer* is the most significant word, whereas *employees* occurs in the top-5 words. *Affordances* are the possible goal-oriented actions by a chatbot for a particular user group. Together, these words signify the usage of chatbots by user roles that are external as well as internal to an *organization*. The words *government* and *enterprise* imply the different types of organizations where chatbots are being integrated, possibly over channels like *Slack*[***]. *Citizen* is another user role in the context of governments. A closer look at the papers related to this topic revealed that a diverse range of application areas like journalism (*news*, *media*), user roles, and strategies (*repair*, *affordances*) are covered in this topic.

Topic 5 was interpreted as studies on the perception of social or behavioral characteristics of chatbots. It includes the words *participant*, *social*, *human*, *research*, *design*, and *study*. The papers related to this topic focus on *design* of chatbots by considering how chatbots are *perceived* by *users* during *social interaction*, how can the interaction be made more *human*-like, and how to improve users' *trust* in chatbots.

Topic 6 was interpreted as chatbots to query information from knowledge bases. We see *question* and *answer* as the most prominent words. Other terms like *query*, *knowledge*, *ontology*, *graph*, *data*, *knowledge_base*, *keyword*, and *question_answer* imply the usage of chatbots in question-answering systems e.g., FAQ chatbots to query information. A knowledge base e.g., ontology or knowledge graph, is used to generate the responses.

Topic 7 has *model* as the most prominent word followed by *response*, *word*, *training*, and *context* and was interpreted as training of neural models for response generation in chatbots. Although some words are common with Topic 1, the terms *seq* and *neural* imply that this topic is specifically about response generation.

Topic 8 was interpreted as user interaction with chatbots. It has *user*, *based*, *information*, *system*, and *use* as the most dominant words. These words are among the words with the highest weights across all topics. Other words like *interaction*, *service*, *question*, *answer*, *response* highlight the focus on interaction.

The interpretation of the final topic, Topic 9 was very clearly as chatbots in healthcare, with *health*, *patient*, *healthcare*, and *mental* as the significant words. Additional terms *mental_health*, *eca*, *disease*, *symptoms*, and *depression* indicate an application of chatbots, particularly Embodied Conversational Agents, in mental health care.

[***] https://slack.com/intl/en-ch/features/channels

5.2 Topics related to Coaching

As described in the methodology, the terms *counseling*, *tutoring*, *mentoring*, and *consultation* were considered synonyms of *coaching*. The term *coaching* appeared in two topics – Topic 3: chatbots in education and Topic 9: chatbots in healthcare, shown in Table 2.

It can be observed that *coaching* has more weight in Topic 3 than Topic 9, implying a slightly more adoption of coaching chatbots in the education domain compared to the healthcare domain. In Topic 3, *coaching* is closely followed by *pedagogical* indicating that coaching (including tutoring and mentoring) and teaching are almost equally important pedagogical methods. A closer look at the documents covered by Topic 9 shows that coaching as counseling is more common in the (mental) healthcare domain.

5.3 Evaluation of Topic Interpretation

To ensure that our topic interpretation was 1. comprehensible to the general reader without referring to the corresponding papers and 2. suitably describing the relevant context and the domain of chatbots, we adopted the "semantic validation through expert evaluation" approach, as prescribed by Asmussen & Møller (2019).

Taking the inter-disciplinary nature of Conversational AI into account, we chose four participants that have expertise in different areas of Conversational AI e.g., Machine Learning, Natural Language Processing (NLP), and application areas like education, social and behavioral aspects, and business applications like customer service. The experts satisfied at least one criterion from a. Professor supervising doctoral theses and/or author of publications on chatbots in top IS venues b. Researcher with experience in research projects on chatbot development c. Young researcher with industry experience in chatbot development. The experts were asked to interpret the topics followed by a brief discussion to understand which words influenced their interpretation. If the interpretation semantically matched our interpretation, the original label was retained. For some topics, the labels were enhanced by including additional terms based on the suggestions of the experts to improve the interpretability. For Topic 5: studies on the perception of social or behavioral characteristics of chatbots and Topic 8: user interaction with chatbots, it was observed that the interpretation varied depending on the area of expertise and the perspective of the expert. The experts assigned more importance to the terms they were familiar with and tended to ignore the less familiar terms while forming the topic label. Also, according to their feedback, the fact that they did not have access to the papers related to a topic had an influence on the interpretation. The original labels for these two topics were derived by referring to the underlying papers and were retained by agreement of the experts.

6 Author Analysis

To answer RQ5, we performed a breakdown of the metadata of the publications and identified the number of publications for every author including co-authorship. The affiliations and the research interests were derived manually from the personal and/or university web page of the authors. For RQ5, our goal was to identify the areas of expertise and how they are distributed internationally.

6.1 Key Authors Active in Conversational AI Research

Figure 5 displays the authors by country and region who have more than five publications in the last five years, either as the sole author or as a co-author. The figure also shows the prominent research areas that have a significant overlap in a particular region.

As can be observed, most authors who have more than five publications in the last five years come from Europe (N=25), out of which most authors come from Germany (N=10). The remaining authors active in Conversational AI research come from the USA, China, and Brazil. The research in Europe originates predominantly from university-affiliated research groups. On the other hand, the authors in the USA, China, and Brazil are all affiliated with leading technology companies like Microsoft, Facebook, and IBM.

The publications by the authors in Figure 5 amount to 5% of the overall publications in the last five years. Around 93% of the publications have been co-authored, and the remaining 7% of

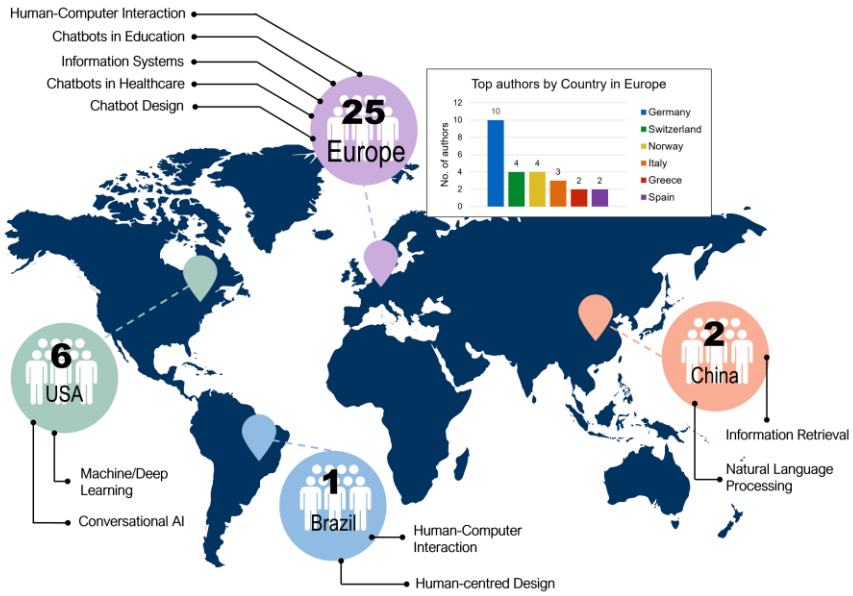

Figure 5: Author and knowledge distribution in Conversational AI publications from 2016-2020

publications have single authors. In Europe, co-authorships can be observed not only between authors from the same research institute or university but also between authors from universities in other European countries, highlighting an overlap in the research interests.

The prominent research areas in the USA are Machine/Deep Learning and Conversational AI. The authors in China have interests in Information Retrieval and Natural Language Processing. In Europe, the significant research areas are Human-Computer Interaction, Information Systems, Chatbots in Education and Healthcare, and Chatbot Design. The design of chatbots encompasses further research interests like persuasive design, anthropomorphic design, and human-centric design. The author in Brazil has an overlap with some authors in Europe in the areas of Human-Computer Interaction and Human-centered Design.

6.2 Key Authors Active in Conversational AI in Coaching

We further identified all papers related to the area of coaching (including synonyms) by manually filtering the papers that discuss coaching and its synonyms and performed a separate author analysis.

Nine authors have more than three publications in the last five years in a domain related to coaching. As also identified in the topic analysis, the dominant coaching domains are healthcare with six authors and education with three authors. The authors come from Switzerland (N=3), Germany (N=2), the UK (N=2), the US (N=1), and Italy (N=1). The healthcare domain has more publications

(N=26) compared to the education domain (N=17). In the healthcare domain, the authors have further specific interests in areas like mental health, personal health and fitness, and mobile health (mHealth). In the education domain, the authors specialize in online and technology-enhanced learning. The papers published by these nine authors (N=43) constitute around 15% of the overall publications related to coaching and about 4% of the total publications in the last five years.

7 Discussion

In the previous sections, we described our analysis to answer the five research questions identified for this work. In this section, we carry out a discussion on our findings and highlight some insights with respect to the research questions.

Looking at the analysis of RQ1 and RQ2, one can assume that the number of publications will continue to grow in the coming years. However, one must also consider the evolution of Gartner's Hype Cycle for Artificial Intelligence for 2020[†††]. Gartner's Hype Cycle provides insights into the expectations of emerging technologies in organizations. Although there is no clear indication of a correlation between the hype cycle and its impact on research, O'Leary (2008) suggests that types of researches in Information Systems, e.g. Design Science Research, Case Study, etc., can be aligned with the position of technology on the Hype Cycle. In 2019, "chatbot" technology was at the peak of inflated expectations[‡‡‡] whereas, in 2020, it slid close to the trough of disillusionment, which according to O'Leary (2008) may discourage research efforts into this particular technology.

The topic analysis results corresponding to RQ3 showed that Topic 8: user interaction with chatbots has the highest cumulative weight followed by Topic 7: training of neural models for response generation in chatbots. Topic 8 highlights the most popular application of chatbots that is to provide information to the users. This is probably due to the fact that chatbots can provide information in an interactive style and find applications in most user-driven scenarios where the user can receive information with only a short-term interaction (Følstad et al., 2019). With respect to Topic 7, it is not surprising to see that several publications focus on training conversational models using neural approaches. Neural and Deep Learning approaches have gained popularity in fields like Information Retrieval, Natural Language Processing, and Machine Learning, which all contribute to the area of Conversational AI (Gao et al., 2019). Moreover, Topic 7 also emphasizes that response generation in chatbots is an emerging area of research.

From our topic analysis, we also found that Topic 4: usage of chatbots in organizations accumulated the lowest cumulative weight. One reason for this could be the high specificity of the topic area. This could also be a result of the configuration of the LDA algorithm. If the algorithm was configured with fewer topics, Topic 4 most likely would have been merged with the other topics. However, despite the lower cumulative weight, this topic highlights some very interesting areas and insights into the state-of-the-art, as described in section 5.1. An additional insight from Topic 6: chatbots to query information from knowledge bases is that knowledge representation is an important aspect of domain-specific conversational agents, e.g., FAQ or Q&A chatbots.

Regarding our analysis for topics related to coaching with respect to RQ4, we observed that the word *coach* or *coaching* did not appear in our top results until we included synonym replacement in the preprocessing steps. The term *coaching* has a relatively lower weight in the two topics where it appears. Our conclusion from this observation is that, with or without synonym replacement, coaching as an application for chatbots is not prominent enough to be captured by topic modeling. Since some chatbot applications in coaching exist in domains like education and healthcare, it seems worthwhile

[†††] https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020/

[‡‡‡] https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019/

to explore the commonalities of coaching as a general application area across different domains. To find specific publications related to coaching like Terblanche (2020) that were not captured in topic modeling, a focused search with keywords representing the characteristics of coaching would be effective. These observations provide us a hint of a research gap where we can channelize our future work.

We also observe that the attribute of emotional behavior is more associated with Embodied Conversational Agents, which is reasonable, as Embodied Conversational Agents can communicate emotions through several conducts like gestures, facial expressions, speech, voice, etc. However, several studies involve emotion detection also in text-based chatbots (Rapp et al., 2021) and apply data labeling as a preferred approach for training empathetic chatbots. But this association between emotion detection and training data was not prominent in any topic, especially in topics that deal with training conversational models i.e., Topic 1 and Topic 7.

From the author analysis for RQ5, one can observe that the research interests of the authors in the USA and China are quite similar with a technical focus. This could be because the authors are affiliated with technology companies. The authors from Europe are focusing mainly on applied research in various areas like education, healthcare, project management and innovation. Referring to the author analysis for coaching, the results would enable us to collaborate with authors that have a similar research interest.

Finally, we would like to mention the limitations of our study. This research was based on two scientific databases and can be extended by including other research databases. However, two databases already provide effective results since DBLP and Scopus also include high-quality publications from IEEE and ACM. The other limitation of our work is linked to the limitation of the LDA algorithm which does not provide a correlation between topics. Using a variant of topic modeling algorithm i.e., Correlated Topic Model (CTM) (Blei & Lafferty, 2005) can provide more information on which topics strongly correlate and provide additional insights. Using another variant like Structural Topic Modeling (STM) (Roberts et al., 2013) allows to also include metadata like author information, keywords, venue, etc., for the topic generation. However, since we wanted to discover the topics based purely on the contents of the publications, we did not include additional metadata in topic modeling and performed the venue and author analysis separately. Lastly, our study does not provide an insight into the evolution of topics over time due to the availability of less volume of data before 2016.

8 Conclusion and Future Work

We carried out a quantitative and qualitative analysis of Conversational AI publications from two scientific sources – DBLP and Scopus. The analysis encompasses different perspectives, like identification of the trend of publications, popular venues and journals, important topics in the publications, and the key authors and their areas of expertise. The trend analysis provides an impression of a growing interest in Conversational AI, and further research in this area can be considered worthful. The analysis of popular venues and journals offers possible research outlets where one can plan to publish research related to Conversational AI as well as refer to the state-of-the-art. The author analysis gives an insight into the key authors, their areas of expertise, their affiliations and possible collaboration opportunities.

The topic analysis presents popular and emerging themes discussed in the literature related to Conversational AI. The identified topics span diverse themes like user interaction, training, use of knowledge bases, behavioral characteristics, and application areas like education and healthcare. The application of conversational agents for coaching-like activities was observed to be prominent in areas of education and healthcare – two areas that have a major contribution to Society 5.0. Empathic

listening is an important coaching tool (Diller et al., 2021), especially desirable in healthcare coaching, and can be realized in conversational agents through emotion-detection and response generation techniques. Coaching conversations have the potential to improve the knowledge, confidence, and decision-making of individuals, thus empowering and positively impacting society. Generating coaching conversations in a conversational agent is, hence, a non-trivial task. Thus, it seems worthwhile to pursue research in Conversational AI in the context of coaching, for which we have identified the main research areas as a. specific characteristics of coaching as an application area of Conversational AI b. emotion-detection in coaching utterances, c. generation of coaching responses and d. knowledge representation of coaching and its domains.

We consider the findings in this work as a starting point for further research in text-based conversational agents which encourage us to identify further strategies related to knowledge representation, emotion detection and response generation in text-based conversational agents.

References

Antons, D., Breidbach, C. F., Joshi, A. M., & Salge, T. O. (2021). Computational Literature Reviews: Method, Algorithms, and Roadmap. *Organizational Research Methods*. https://doi.org/10.1177/1094428121991230

Asmussen, C. B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, *6*, 93. https://doi.org/10.1186/s40537-019-0255-7

Bavaresco, R., Silveira, D., Reis, E., Barbosa, J., Righi, R., Costa, C., Antunes, R., Gomes, M., Gatti, C., Vanzin, M., Junior, S. C., Silva, E., & Moreira, C. (2020). Conversational agents in business: A systematic literature review and future research directions. *Computer Science Review*, *36*, 100239. https://doi.org/10.1016/j.cosrev.2020.100239

Bernardini, A. A., Sônego, A. A., & Pozzebon, E. (2018). Chatbots: An Analysis of the State of Art of Literature. *Anais Do I Workshop on Advanced Virtual Environments and Education (WAVE 2018)*, *1*, 1–6. https://doi.org/10.5753/wave.2018.1

Blei, D. M., & Lafferty, J. D. (2005). Correlated topic models. *Advances in Neural Information Processing Systems*, *18*, 147–154. https://dl.acm.org/doi/10.5555/2976248.2976267

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://dl.acm.org/doi/10.5555/944919.944937

Caldarini, G., Jaf, S., & McGarry, K. (2022). A Literature Survey of Recent Advances in Chatbots. *Information (Switzerland)*, *13*(1), 41. https://doi.org/10.3390/info13010041

Calvaresi, D., Ibrahim, A., Calbimonte, J.-P., Schegg, R., Fragniere, E., & Schumacher, M. (2021). The Evolution of Chatbots in Tourism: A Systematic Literature Review. *Information and Communication Technologies in Tourism 2021*. https://doi.org/10.1007/978-3-030-65785-7_1

Cavacini, A. (2015). What is the best database for computer science journal articles? *Scientometrics*, *102*, 2059–2071. https://doi.org/10.1007/s11192-014-1506-1

Darling, W. M. (2011). A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1–10.

De Lacerda, A. R. T., & Aguiar, C. S. R. (2019). FLOSS FAQ chatbot project reuse - How to allow nonexperts to develop a chatbot. *Proceedings of the 15th International Symposium on Open Collaboration, OpenSym 2019*, 1–8. https://doi.org/10.1145/3306446.3340823

Diller, S. J., Mühlberger, C., Löhlau, N., & Jonas, E. (2021). How to show empathy as a coach: The effects of coaches' imagine-self versus imagine-other empathy on the client's self-change and

coaching outcome. *Current Psychology*. https://doi.org/10.1007/s12144-021-02430-y

Fletcher, S. J., & Mullen, C. A. (2012). The SAGE handbook of mentoring and coaching in education. In *The SAGE Handbook of Mentoring and Coaching in Education*. SAGE Publications Ltd. https://doi.org/10.4135/9781446247549

Følstad, A., & Skjuve, M. (2019). Chatbots for customer service: User experience and motivation. *ACM International Conference Proceeding Series*, 1–9. https://doi.org/10.1145/3342775.3342784

Følstad, A., Skjuve, M., & Brandtzaeg, P. B. (2019). Different chatbots for different purposes: Towards a typology of chatbots to understand interaction design. *Internet Science - {INSCI} 2018 International Workshops*, *11551 LNCS*, 145–156. https://doi.org/10.1007/978-3-030-17705-8_13

Gale, L. (2017). Sport coaching concepts: a framework for coaching practice (2nd ed). *Sports Coaching Review*, 212–216. https://doi.org/10.1080/21640629.2017.1409965

Gao, J., Galley, M., & Li, L. (2019). Neural Approaches to Conversational AI. *Foundations and Trends® in Information Retrieval*, *13*, 127–298. https://doi.org/10.1561/1500000074

Grant, A. M., Passmore, J., Cavanagh, M. J., & Parker, H. (2010). The State of Play in Coaching Today: A Comprehensive Review of the Field. In G. P. H. & J. K. Ford (Ed.), *International review of industrial and organizational psychology* (pp. 125–167). Wiley. https://doi.org/10.1002/9780470661628.ch4

Härer, F., & Fill, H.-G. (2020). Past Trends and Future Prospects in Conceptual Modeling-A Bibliometric Analysis. *International Conference on Conceptual Modeling*, 34–47. https://doi.org/10.1007/978-3-030-62522-1_3

Hussain, S., Ameri Sianaki, O., & Ababneh, N. (2019). A Survey on Conversational Agents/Chatbots Classification and Design Techniques. In L. Barolli, M. Takizawa, F. Xhafa, & T. Enokido (Eds.), *Advances in Intelligent Systems and Computing* (Vol. 927, pp. 946–956). Springer International Publishing. https://doi.org/10.1007/978-3-030-15035-8_93

Io, H. N., & Lee, C. B. (2017). Chatbots and conversational agents: A bibliometric analysis. *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, *2017-Decem*, 215–219. https://doi.org/10.1109/IEEM.2017.8289883

Irby, B. J., Boswell, J., Jeong, S., & Pugliese, E. (2018). Editor's overview: tutoring and coaching. *Mentoring & Tutoring: Partnership in Learning*, *26*, 245–248. https://doi.org/10.1080/13611267.2018.1511955

Jurafsky, D., & Martin, J. H. (2020). Chatbots and Dialogue Systems. In D. Jurafsky & J. H. Martin (Eds.), *Speech and Language Processing.* (Draft 3rd, p. Chapter 24). https://doi.org/https://web.stanford.edu/~jurafsky/slp3/24.pdf

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., & Theeramunkong, T. (2019). A survey on evaluation methods for chatbots. *ACM International Conference Proceeding Series, Part F1483*, 111–119. https://doi.org/10.1145/3323771.3323824

Muff, F., Härer, F., & Fill, H.-G. (2022). Trends in Academic and Industrial Research on Business Process Management - A Computational Literature Analysis. *Proceedings of the 55th Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2022.874

O'Leary, D. E. (2008). Gartner's hype cycle and information system research issues. *International Journal of Accounting Information Systems*, *9*, 240–252. https://doi.org/10.1016/j.accinf.2008.09.001

Okoli, C. (2015). A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*, *37*(1), 879–910. https://doi.org/10.17705/1CAIS.03743

Pamungkas, E. W. (2019). Emotionally-aware chatbots: A survey. *ArXiv Preprint ArXiv:1906.09774*. https://doi.org/10.48550/arXiv.1906.09774

Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education:

A systematic literature review. *Computer Applications in Engineering Education*, *28*(6), 1549–1565. https://doi.org/10.1002/cae.22326

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

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. *NIPS 2013 Workshop on Topic Models*, 2–5.

Schofield, A., & Mimno, D. (2016). Comparing Apples to Apple: The Effects of Stemmers on Topic Models. *Transactions of the Association for Computational Linguistics*, *4*, 287–300. https://doi.org/10.1162/tacl_a_00099

Schröder, G., Thiele, M., & Lehner, W. (2011). Setting goals and choosing metrics for recommender system evaluations. *CEUR Workshop Proceedings*, *811*, 78–85. https://www.researchgate.net/publication/268381252

Skjuve, M., & Brandzaeg, P. B. (2019). Measuring User Experience in Chatbots: An Approach to Interpersonal Communication Competence. In *International Conference on Internet Science* (pp. 113–120). Springer. https://doi.org/10.1007/978-3-030-17705-8_10

Sutton, C., & Gong, L. (2017). Popularity of arXiv.org within Computer Science. *ArXiv Preprint ArXiv:1710.05225*.

Terblanche, N. (2020). A design framework to create artificial intelligence coaches. *International Journal of Evidence Based Coaching and Mentoring*, *18*(2), 152–165. https://doi.org/10.24384/b7gs-3h05

Wolever, R. Q., Moore, M. A., & Jordan, M. (2017). Coaching in healthcare. *The Sage Handbook of Coaching*, 521–543.