Artificial Intelligence in Customer Service Interactions

From Multi-Layered Information to Organizational Insights

Bea Waelbers

ISBN: 978-94-6510-999-2

Cover design & lay-out: herikmedia
Printed by: proefschriftmaken.nl

Artificial Intelligence in Customer Service Interactions

From Multi-Layered Information to Organizational Insights

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Open Universiteit
op gezag van de rector magnificus
prof. dr. Th.J. Bastiaens
ten overstaan van een door het
College voor promoties ingestelde commissie
in het openbaar te verdedigen

op vrijdag 19 december 2025 te Heerlen
om 13.30 uur precies

door
Bea Maria Lucia Waelbers
geboren op 25 april 1997 te Neerpelt, België

Promotores:
Prof. dr. S. Bromuri, Open Universiteit
Prof. dr. H.P. van Ditmarsch, University of Toulouse

Copromotor:
Dr. A.P. Henkel, Open Universiteit

Leden beoordelingscommissie:
Prof. dr. J.G.A.M. Lemmink, Maastricht University
Prof. dr. F. Toni, Imperial College London
Prof. dr. N.A. Alechina, Open Universiteit
Prof. dr. P.L. Curşeu, Open Universiteit

Table of Contents

Chapter 1

Introduction

1.1 The role of service interactions in the service economy

Services are at the heart of how we live, work, and connect in today's world. Customers engage daily with service providers to order meals, stream movies, arrange deliveries, access healthcare, and resolve technical issues. These everyday interactions demonstrate the significance of service in the global economy, accounting for up to 78% of Western economic activity and 66% worldwide (O'Neill, 2025; World Bank, 2024). As the service sector continues to expand, it reshapes both consumer behavior and business strategies. The growing emphasis on services shows the changing expectations of customers, who increasingly demand seamless experiences, convenience, and personalization (Ameen, Tarhini, Reppel, & Anand, 2011). For businesses, this shift necessitates a strategic focus on service quality as a key differentiator. Companies that excel in customer service not only gain a competitive edge but also lay the foundation for sustained success and customer loyalty (Sheth, Jain, & Ambika, 2023; Tahir, Adnan, & Saeed, 2024).

As these firm-customer interactions have increased substantially in the last decades, consumers now engage in direct contact with companies more frequently than ever before. This growing volume of interactions reflects a broader shift from goods-centered exchanges toward a service-based economy, where co-production, personalization, and continuous adaptation have become central mechanisms of competitive advantage (Vargo & Lusch, 2008). To meet customers' diverse needs within this new context, companies have incorporated a wide variety of communication channels (Gao, Fan, Li, & Wang, 2021). What once consisted primarily of telephone support and in-person interactions has evolved into a complex infrastructure encompassing emails, live chats, social media, mobile applications, and video calls. Organizations have restructured their support systems accordingly, employing millions of customer service agents globally (Bohne, Raphael, 2024). This workforce represents a critical interface between organizations and their customers, handling everything from basic inquiries and technical support to complex problem resolution and sales activities (Aksin, Armony, & Mehrotra, 2009).

In these service interactions, value is co-created through information exchange between customers and service providers (Grönroos, 2011; Vargo & Lusch, 2008). Co-creation refers to the idea that the quality and outcome of a service interaction

are shaped jointly by the actions of both the customer and the provider during the interaction (Vargo & Lusch, 2008). These co-creation opportunities are central for firms because they are the points at which customer expectations, emotions, and company actions meet, strongly influencing perceptions of quality and satisfaction (Fuentes, Smyth, & Davies, 2019; van Dolen, Lemmink, de Ruyter, & de Jong, 2002).

However, while service-dominant logic and related frameworks explain who participates in value co-creation, they under-theorize how the informational complexity of these interactions can be systematically processed and used for organizational learning (Grönroos, 2011; Vargo & Lusch, 2008). While this relational perspective has enhanced insights into customer-firm interactions, it fails to conceptualize service interactions as autonomous information systems. Consequently, the informational complexity of service interactions and the organizational challenge of processing this information remain underdeveloped in service theory (Bardhan, Demirkan, Kannan, Kauffman, & Sougstad, 2010; Marinova, de Ruyter, Huang, Meuter, & Challagalla, 2016).

1.2 The information potential in service interactions

While customer interactions generate enormous amounts of potentially valuable data, effectively utilizing this information presents significant challenges. The volume and complexity of these data overwhelm human cognitive processing capabilities, making it demanding for service agents to process and extract actionable insights (Miller, 1956). This difficulty is amplified by the fact that service agents already operate in high-stress environments, particularly given the emotional demands of their work (Aksin et al., 2009; Grandey, Dickter, & Sin, 2004). Service agents must manage not only customer emotions but also their own responses, which can be particularly taxing when dealing with frustrated or distressed customers (Grandey et al., 2004). Additionally, they must balance multiple responsibilities such as meeting performance metrics, managing response times, and navigating complex technological systems (Tovar, 2021). These metrics often create competing demands, as service agents must balance efficiency with quality, speed with thoroughness, and problem-solving with sales objectives. These competing demands add to their stress levels and can lead to burnout and high turnover rates. This then results in substantial costs for organizations in terms of recruitment, training, and lost productivity (T.-Y. Park &

Shaw, 2013; Zito et al., 2018).

These operational and human pressures limit service agents' capacity to capture and communicate key information during and after interactions accurately. As a result, a large portion of valuable data goes unprocessed and cannot contribute to organizational learning (Kumar et al., 2013). Despite their importance, current service theory provides limited guidance on how firms can systematically process and learn from customer service interactions. While co-creation has advanced our understanding of the relational nature of services by emphasizing the role of multiple stakeholders in jointly shaping value, it offers little insight into how the informational outcomes of these interactions can be analyzed or operationalized (Grönroos, 2011). It highlights who is involved in creating value, but not how the information exchanged, often embedded in dialogue, emotion, and behavior, can be extracted and used. Moreover, current frameworks provide limited guidance on how service firms can systematically process the complex, unstructured conversational data that characterizes today's high-volume and multimodal service contexts. This gap has become increasingly critical as organizations generate thousands of daily interactions across multiple channels, yet lack coherent frameworks for understanding how to extract and utilize the rich information embedded within these conversations at scale. Given the large scale of modern customer service operations, each customer interaction simultaneously poses the challenge of delivering optimal support while providing valuable opportunities to collect detailed information about customer preferences, problems, and behaviors to aid organizational learning (Henkel, Bromuri, Iren, & Urovi, 2020; Prentice, Lopes, & Wang, 2019).

Despite this rich information potential, current analysis methods capture only a fraction of what is available. Organizations often struggle to convert the tacit knowledge embedded in conversations into explicit, actionable insights (H. Chen, Chiang, & Storey, 2012). While some organizations have started automating parts of conversation analysis, many still rely on established approaches such as post-interaction surveys and manual quality monitoring of small samples (McKinsey & Company, 2024). These methods, although useful, are often resource-intensive and typically provide delayed insights, limiting their effectiveness for real-time decision-making.

Service interactions are often treated as discrete, functional exchanges, rather than as complex events that generate multiple layers of information (Keith, Lee, & Gravois Leem, 2004). In reality, however, these interactions simultaneously involve explicit content (e.g., problem descriptions or solutions), implicit signals (e.g., emo-

tion, urgency, intent), and broader patterns that emerge across interactions (Henkel, Bromuri, et al., 2020; Papadia, Pacella, Perrone, & Giliberti, 2023; Tong, Jia, Luo, & Fang, 2021). To address the theoretical gap identified above, this dissertation reconceptualizes service interactions as multi-layered information structures. This framing extends service theory by highlighting that service interactions are not only sites of co-creation but also rich information systems that must be systematically processed. This shift enables service theory to move beyond its current limitations and better account for the complexities of modern service interactions.

This perspective allows us to conceptualize service conversations as containing multiple layers of information, with at least three distinct information layers. First, interactions can be studied at a content level, where information is extracted from the exact words spoken (Papadia et al., 2023). Here, one can identify customer problems, product details, and resolution steps. This explicit content represents the most accessible layer of information available in service conversations.

Second, there is an implicit information layer, where underlying information can be extracted beyond the literal content. This includes the complexity of the problem and its urgency, which can be inferred from language patterns and communication style (Zolfagharian, Hasan, & Iyer, 2018). Emotional information can also be extracted at this level, revealing customer satisfaction, frustration, and other emotional states throughout the interaction (Henkel, Bromuri, et al., 2020). Behavioral patterns emerge from both customers and service agents, showing how they navigate conversations and approach problem-solving (Van Herck, Decock, & Fastrich, 2022).

Third, customer interactions can also provide insights at an aggregated level. Instead of focusing only on single interactions, analyzing patterns across many conversations makes it possible to detect recurring issues, evaluate how service processes function as a whole, and identify systematic strengths and weaknesses in the way organizations handle customer requests (Van Herck et al., 2022). Such an aggregated perspective can inform training priorities, process redesign, and the development of organizational best practices and is aimed to improve overall service quality (Jarvenpaa & Välikangas, 2025).

When effectively extracted and utilized, these multi-layered information structures can reduce cognitive and emotional load on service agents as they allow for timely insights and guidance, allowing for more efficient responses (H. Chen et al., 2012). Moreover, systematic extraction and analysis of the rich, multi-layered information contained in service interactions can foster organizational learning at an unprece-

dented scale, enhancing both tactical and strategic decision-making (Bardhan et al., 2010; Cohen, 2018; Kumar et al., 2013). This enables firms to identify recurring pain points, recognize shifts in customer sentiment, personalize service delivery, and translate these insights into concrete improvements in processes, training, and support tools (Choi, 2018; Kumar et al., 2013).

To bridge this gap between the information potential of the multi-layered conversations and actual extraction, advances in artificial intelligence (AI) offer promising solutions. AI provides the computational power and pattern recognition capabilities to process large volumes of conversational data automatically and consistently (Goodfellow, Bengio, & Courville, 2016; Russell & Norvig, 2016). This systematic analysis of unstructured, multi-layered conversational data supports practical information extraction while it advances information systems theory on organizational information processing capabilities (Balducci & Marinova, 2018; H. Chen et al., 2012). This capability facilitates organizational learning at a large scale as it captures knowledge embedded in service interactions that would otherwise be lost, thereby deepening the theoretical understanding of how organizations learn from customer interactions.

1.3 Artificial intelligence in customer service

The challenge of extracting actionable insights from multi-layered service interactions requires reconceptualizing how organizations process information from customer interactions. This theoretical gap calls for computational approaches that can systematically decode the complex information structures embedded in service conversations. Artificial intelligence represents a domain of computer science focused on developing intelligent algorithms that can perform tasks traditionally requiring human cognition (McCarthy, Minsky, Rochester, & Shannon, 1955; Russell & Norvig, 2016). At its foundation are neural networks, which mimic biological neurons as interconnected nodes that process and transmit information (Janiesch, Zschech, & Heinrich, 2021). Building on this foundation, deep learning utilizes multi-layered neural networks to discover patterns in datasets automatically (LeCun, Bengio, & Hinton, 2015). These advances have enabled the development of complex applications such as neural machine translation systems and large language models. Large language models, in particular, are trained on vast datasets and can generate human-like responses across a wide range of tasks, including comprehending context, performing

reasoning, and generating effective responses (Radford, Narasimhan, Salimans, & Sutskever, 2018; Vaswani et al., 2017).

The fundamental capability underlying these AI technologies is pattern recognition, which enables algorithms to identify regularities, anomalies, and relevant structures within data. In the context of customer service, AI's implementation for data-driven service management addresses the theoretical challenge of converting tacit knowledge embedded in multi-layered service conversations into explicit organizational assets (Kumar et al., 2013; Rust & Huang, 2014). This examination of AI model effectiveness for different service information extraction types enables researchers to infer fundamental properties about service communication patterns and embedded information. For instance, if deep learning models significantly outperform simpler approaches, this suggests that service conversations contain complex, non-linear information structures that simpler models do not capture (Goodfellow et al., 2016).

The challenge of extracting actionable insights from multi-layered service interactions requires the reconceptualization of how organizations process information from customer interactions. In this dissertation, artificial intelligence is not treated merely as a technical tool, but as a theoretical instrument that enables service firms to process multi-layered information structures at scale. In doing so, AI operationalizes the idea that service interactions function as information systems, thereby advancing service theory and information systems theory by establishing computational information processing as a core organizational capability.

From a management science perspective, this dual functionality of AI represents a shift toward the use of data-driven service management, where AI is not simply a practical tool for data processing but also a theoretical instrument that helps decode the characteristics of service interactions (Maglio & Spohrer, 2008). Expanding on this viewpoint, the theoretical foundation for AI-driven service analytics is that the systematic extraction of information can convert implicit knowledge from service interactions into explicit organizational assets (Choi, 2018). This transformation is particularly significant in service contexts, where much of the valuable information exists in conversational interactions that have traditionally been difficult to analyze systematically. Through its ability to process natural language and identify patterns in human communication, AI creates new possibilities for organizational learning from service interactions (Marinova et al., 2016).

Building on information processing theory and knowledge management frameworks, service firms can utilize AI's pattern recognition capabilities to develop sys-

tems that systematically extract and operationalize the multi-layered information embedded in service interactions (Choi, 2018). Automated extraction reduces the cognitive burden on service agents while enhancing their awareness of customer needs, emotional states, and conversational dynamics, and can further support their development through performance feedback (A. Ahmed, Shaalan, Toral, & Hifny, 2021; Bromuri, Henkel, Iren, & Urovi, 2021; Henkel, Bromuri, et al., 2020; Tong et al., 2021). Given the scale and complexity of modern customer service operations, such automation is essential, as manual analysis would be impractical. Research has explored how these capabilities influence multiple dimensions of the customer service process. Natural language processing has enabled analysis of customer communications (Shah, Ghomeshi, Vakaj, Cooper, & Fouad, 2023). Emotion recognition technologies have identified emotional cues (Henkel, Bromuri, et al., 2020). AI's predictive capabilities have been applied to customer churn prediction and loyalty analysis (Prentice et al., 2019). Service agent assist tools provided real-time support, suggesting responses and guiding service agents through complex scenarios (Bromuri et al., 2021; Dong & Srinivasan, 2013). Together, these applications illustrate how computational tools can systematically process multi-layered service information, advancing organizational information processing capabilities in service contexts.

Moreover, from a computer science perspective, customer service represents a valuable application domain for developing and refining intelligent systems. Customer service interactions are inherently multimodal, encompassing voice signals, textual exchanges, and even emotional cues, all of which pose substantial challenges for AI systems to interpret cohesively (Bardhan et al., 2010). This requires the integration of natural language processing, speech recognition, sentiment analysis, and machine learning, pushing the boundaries of current AI architectures. As such, studying customer service data contributes directly to the field of computer science, fostering the development of models capable of handling heterogeneous data sources, addressing real-world noise and ambiguity, and performing in dynamic, high-stakes environments (Choi, 2018). The standardized yet variable nature of customer service interactions, with predictable structures but diverse content, makes them particularly well-suited for AI analysis.

The systematic application of AI to service analytics requires theoretical frameworks that bridge computational capabilities with organizational information needs. The CRISP-DM (Cross Industry Standard Process for Data Mining) framework provides a meta-theoretical structure for understanding how AI systems can be designed

to extract and operationalize service intelligence (Martínez-Plumed et al., 2021). Developed for industrial data mining, it provides a systematic approach for connecting domain-specific service challenges with the development of computational models. The framework consists of phases, including business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This sequence reflects how AI systems are designed in close relation to organizational objectives and the characteristics of empirical data. CRISP-DM supports integrating these diverse data sources of customer interactions into a unified modeling process (Marinova et al., 2016). Its iterative nature allows findings from one phase to guide adjustments in others, which is essential when interpreting complex conversational data for applications like service agent support or customer feedback analysis (A. Ahmed et al., 2021). Beyond practical utility, CRISP-DM contributes to information systems theory by formalizing the cyclical relationship between theoretical inquiry and empirical modeling in organizational contexts. It enables systematic exploration of how AI can represent and reason about rich, multimodal service data, advancing understanding of both AI methodologies and the computational requirements for processing service information.

However, while significant progress has been made in applying AI to service management, there remains a theoretical gap in understanding how different computational approaches can systematically extract the multi-layered information embedded in service interactions. This gap represents an opportunity to advance knowledge at the intersection of information systems theory, artificial intelligence, and service management (Bardhan et al., 2010; Borges, Laurindo, Spínola, Gonçalves, & Mattos, 2021; Choi, 2018). Rather than treating AI as a standalone innovation, it should be conceptualized as a mechanism through which service organizations can develop the computational information processing capabilities needed to engage with multi-layered service interactions systematically. The systematic evaluation of different AI models for service information extraction can reveal fundamental properties about service data characteristics and communication patterns, contributing to the theoretical development of computational information processing in service management (Larivière et al., 2017). This research advances service theory, establishing computational information processing as a core organizational capability for modern service firms, and information systems theory, demonstrating how AI can systematically extract multi-layered intelligence from unstructured organizational interactions.

1.4 Research questions and objectives

This dissertation advances the theory of service firm information processing, demonstrating how AI can systematically extract different types of information from the multi-layered structures embedded within customer service conversations. Figure 1.1 illustrates this conceptual framework, mapping the complete pipeline from customer-service agent interactions to actionable business insights. The framework integrates the three information layers established with AI processing capabilities. It thereby demonstrates how these AI approaches systematically extract information from each conversational layer of customer service interactions, ultimately converting multi-layered service interactions into theoretical insights and strategic business outcomes.

The complexity and scale of modern service operations, combined with the rich, multi-layered information embedded within customer-agent interactions, create both an opportunity and a necessity for systematic AI-driven analysis, as illustrated in the framework. This research contributes to the growing body of knowledge on human-AI collaboration in service contexts. It provides systematic methods for analyzing content, implicit, and aggregated layers of service conversations, benefiting both customers and service agents.

To investigate the potential of artificial intelligence in this context, this dissertation addresses the following main research question:

How can artificial intelligence be employed to extract conversational patterns and information from customer service interactions?

To comprehensively address this main research question, this dissertation investigates three interconnected dimensions. The first step in implementing AI-driven information extraction is selecting appropriate methods that can effectively process and interpret the specific data input. Customer service interactions offer opportunities for applying techniques such as speech emotion recognition, quality assessment, and identifying effective response strategies. Different AI methodologies, ranging from simple multi-layer perceptrons and neural networks to large language models, offer various advantages and limitations when applied to extraction tasks in service conversations. Understanding which AI approaches work best for specific information extraction challenges is crucial for developing robust and reliable systems that can augment human service agents. This leads to the first subquestion:

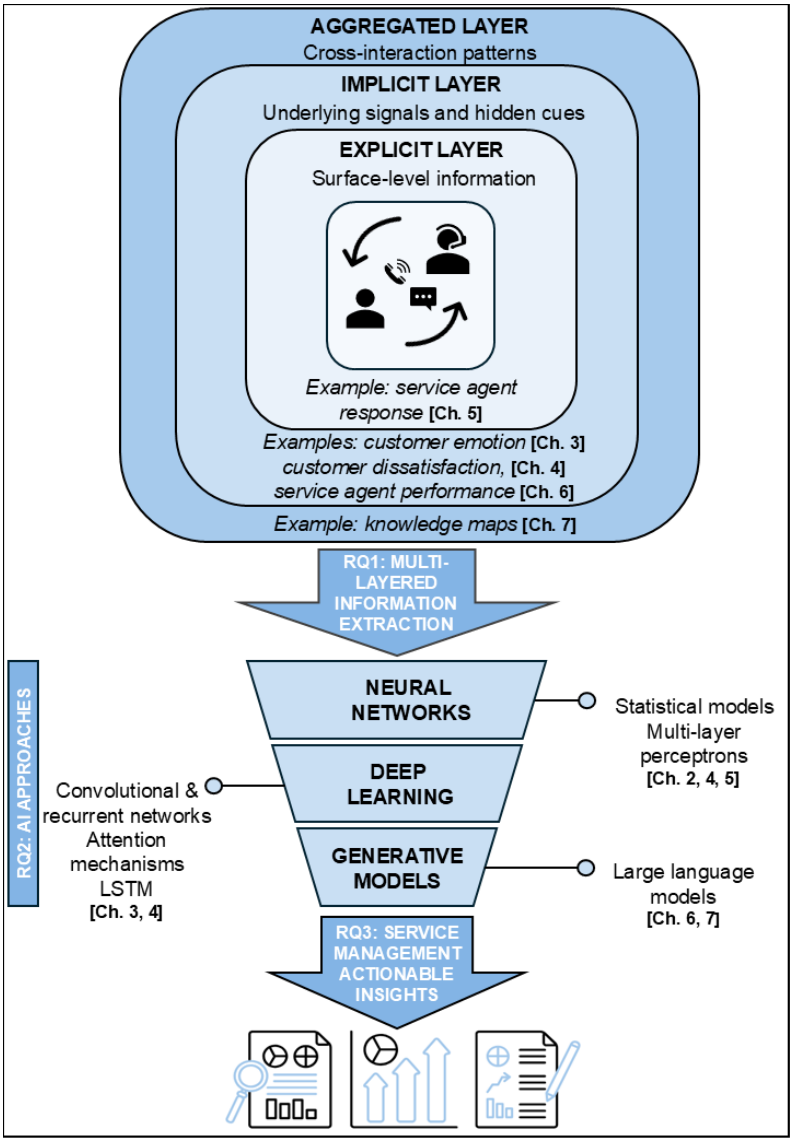

Figure 1.1: Conceptual framework

Note. This figure presents the conceptual framework of the dissertation. It illustrates how insights can be derived from customer interactions across multiple channels. The framework proposes a multi-layered information structure: (1) explicit information, (2) implicit information, and (3) aggregated information. Various AI techniques are employed to extract insights into service management.

Research Question 1: What artificial intelligence approaches can effectively extract information from customer service interactions?

In addition to selecting appropriate AI approaches, it is essential to understand how these technologies can analyze the structure and contextual elements of customer conversations to identify patterns across content, implicit, and aggregated information layers. As customer service interactions contain multiple layers of information, each model requires different analytical capabilities to extract these specific patterns. This analytical capability enables organizations to move beyond simple classification toward a nuanced understanding of the interaction. Following this, the second subquestion is:

Research Question 2: How can artificial intelligence be used to analyze conversational structures and identify meaningful patterns in customer service interactions?

The ultimate value of AI-driven information extraction lies in its ability to generate actionable insights that enhance understanding of service interactions and support human-AI collaboration in service delivery. While technical capabilities and analytical methods are important, organizations must develop a deep understanding of their data characteristics and of the methodologies required to extract relevant insights from customer service conversations. A clear view of the strengths and limitations of different AI approaches, and of how their outputs align with human expertise, is essential for transforming extracted patterns into substantial intelligence. These considerations highlight the need to understand how computational information processing can enhance organizational capabilities in service contexts. This results in the third subquestion:

Research Question 3: What insights can artificial intelligence-driven information extraction provide to service firms for understanding service interactions?

Together, these research questions provide a comprehensive framework for investigating AI's potential to systematically extract patterns from customer service conversations, advancing both theoretical understanding and practical applications in service analytics.

1.5 Dissertation outline

This dissertation comprises six studies: a systematic literature review in Chapter 2, followed by five empirical studies spanning Chapters 3 through 7, and a concluding discussion in Chapter 8. Figure 1.2 presents the visual outline of the dissertation.

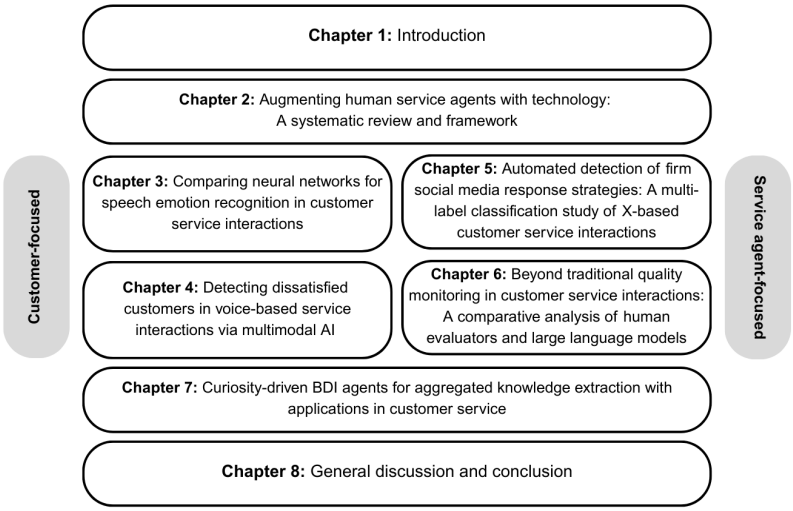

Figure 1.2: Outline of the dissertation

Chapter 2 discusses how core technological capabilities can effectively augment human service agents through a systematic review of 99 empirical studies. The study reveals six central themes of technology integration in service interactions, and proposes a conceptual framework anchored in socio-technical systems theory. This chapter structures existing AI approaches and their effectiveness in service contexts, thus addressing Research Question 1. It also contributes to Research Question 3 by identifying critical aspects of human-AI collaboration that organizations must consider for effective service delivery implementation. This chapter is based on a manuscript currently under revision at the second round at the *Journal of Service Management*.

Chapter 3 compares different neural network architectures for automatic speech emotion recognition in customer service conversations. Here, multi-layer percep-

trons, convolutional neural networks, and neural machine translation models are evaluated against baseline classifiers classifying discrete emotions from speech features from 363 Dutch call center interactions. This chapter uses direct comparison of AI approaches to address Research Question 1, assessing their effectiveness in extracting emotional information from customer service interactions. It is also connected to Research Question 2, as it demonstrates how AI can analyze conversational structures to identify emotional patterns in voice-based service communications. This chapter is based on an article published as Waelbers, B., Bromuri, S., & Henkel, A. P. (2022, July). Comparing neural networks for speech emotion recognition in customer service interactions. *In 2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Chapter 4 investigates machine learning techniques for predicting customer dissatisfaction from voice-to-voice service interactions. The study analyzes 1,144 service interactions from a global service provider to demonstrate how verbal and vocal components should be optimally integrated for dissatisfaction detection. Grounded in communication theory's interactivity principle, the chapter suggests cross-attention as the optimal mechanism to combine the different modalities. This chapter addresses Research Question 1 through evaluating and comparing machine learning approaches for extracting meaningful insights from customer service interactions, explicitly focusing on dissatisfaction prediction capabilities. It connects to Research Question 2 by examining how AI can analyze conversational structures through multimodal integration of verbal and vocal signals to identify complex dissatisfaction patterns. This chapter contributes to Research Question 3 with practical implications for service firms about combining modalities when measuring customer dissatisfaction. This chapter is based on a manuscript currently under revision at the first round at the *Journal of Interactive Marketing*.

Chapter 5 examines automated detection of firm social media response strategies in X-based customer service interactions. It compares large language models with deep learning approaches to extract response strategies from 5,299 consumer-firm interactions. This chapter relates to Research Question 1 by evaluating the effectiveness of different AI approaches in extracting strategic information from social media service interactions. It also examines how AI extracts strategic patterns within single social media interactions, addressing Research Question 2. For Research Question 3, it shows how AI-driven extraction gives companies actionable insights into their response strategies and helps optimize their social media customer service. This chap-

ter is based on an article accepted as Waelbers, B., Henkel, A. P., & Bromuri, S. (in press). Automated detection of firm social media response strategies: A multi-label classification study of X-based customer service interactions. In *10th International Conference on Machine Learning Technologies* (pp. 310-315). IEEE.

Chapter 6 explores the potential of large language models to support human evaluators during call center quality monitoring processes. It investigates how large language models can identify cases where human evaluators may have made errors in quality monitoring forms through a three-step comparative analysis approach from 244 recordings of customer interactions. This chapter contributes to Research Question 1 by assessing large language models as viable AI methodologies for quality assessment. This chapter reveals critical insights about when AI successfully complements human judgment versus when it falls short, addressing Research Question 3. This helps organizations to develop a deeper understanding of human-AI collaboration dynamics in monitoring workflows. This chapter is based on an article published as Waelbers, B., Henkel, A. P., & Bromuri, S. (2026). Beyond traditional quality monitoring in call centers: A comparative analysis of human evaluators and large language models. In: Li, S. (eds) Information Management. ICIM 2025. Communications in Computer and Information Science, Vol. 2540. (pp.1-10). Springer, Cham. https://doi.org/10.1007/978-3-031-99353-4_27.

Chapter 7 proposes a belief-desire-intention (BDI) agent as a curiosity-driven methodology for ontology generation in knowledge extraction tasks. The approach uses large language models to dynamically generate questions guiding document retrieval, outperforming traditional methods on benchmark datasets. While most chapters in this dissertation examine how AI extracts patterns from real-time service conversations, organizations also rely on extensive background knowledge, from training manuals and troubleshooting guides to policy documents and quality monitoring forms. This study suggests that the BDI-driven ontology generation methodology could be applied to these organizational knowledge resources, dynamically structuring them to connect conversational insights with broader firm-level knowledge. In doing so, it extends the dissertation's multi-layered perspective from frontline interactions to organizational learning. This chapter introduces a novel AI methodology that moves beyond passive reasoning, advancing Research Question 1. The work shows how AI can adaptively explore document structures to generate knowledge patterns, supporting Research Question 2. Organizations gain insights into dynamic versus static approaches for scalable knowledge extraction, informing Research Question

3. This chapter is based on a manuscript currently under preparation for submission to the *Symposium of Applied Computing*.

Chapter 8 synthesizes the findings from the preceding studies, highlights their theoretical and practical contributions, and situates them within broader debates on service management and AI-driven information processing. It also addresses ethical considerations, research limitations, and avenues for future work, providing a cohesive conclusion to the dissertation.

Chapter 2

Augmenting human service agents with technology: A systematic review and framework

This chapter is based on a manuscript currently under revision (second round) at the *Journal of Service Management* in collaboration with Dr. Alexander P. Henkel and Prof. Dr. Stefano Bromuri.

Abstract

The integration of advanced technologies into customer service has raised critical questions about how these tools can best augment the roles of human service agents. This literature review documents how core technological capabilities, including artificial intelligence, machine learning, deep learning, and natural language processing, can effectively augment, rather than replace, service agents across the various stages and tasks in customer service interactions. Grounded in socio-technical systems theory, this study provides a systematic and cross-disciplinary review of 99 empirical studies on applications of technology in human-to-human service interactions published over the last decade. Thematic mapping of these studies reveals six central themes: pre-service optimization, interaction intelligence, service agent well-being, service agent monitoring, emotion work, and augmentation. These themes are combined into a conceptual framework, from which a comprehensive agenda for future research is developed. The findings underscore the need for a human-centered approach, demonstrating how technology can enhance service agent roles while promoting well-being and fostering collaboration in customer service. This review shifts the conversation from technology replacing service agents to technology supporting service agents in their roles, contributing to an understanding of how technology can augment service agent effectiveness, efficiency, and well-being, ultimately enabling sustainable improvements in service interactions.

2.1 Introduction

Every day, frontline service agents shape billions of customer experiences. In 2023, over 2.8 million service agents in the US managed customer inquiries in contact centers through email, live chat, and phone (Bohne, Raphael, 2024), not counting the many more working in face-to-face environments. Service interactions form the foundation of value co-creation (Grönroos, 2011; Vargo & Lusch, 2008), but the nature of these interactions is changing rapidly. Advances in technology, including artificial intelligence (AI), analytics, and digital tools, are fundamentally changing how service agents work and the expectations placed upon them (M.-H. Huang & Rust, 2018; Larivière et al., 2017).

The dominant focus in the service literature and industry has been on automation, where systems replace human labor, streamline operations, or deliver self-service experiences (Bowen, 2016; Chi, Denton, & Gursoy, 2020). However, while 67% of customer service leaders have implemented AI in their customer service processes (IBM, 2024), service firms maintain reliance on human service agents (Bowen, 2024), and consumers often prefer or expect human contact, exhibiting aversion towards algorithm-provided service (Ameen et al., 2011; Bowen, 2024; Castelo, Bos, & Lehmann, 2019). Consequently, there is increasing attention to how technology can augment, rather than replace, human service work, enabling service agents to access information, understand customer needs, and manage complex interactions (De Keyser, Köcher, Alkire, Verbeeck, & Kandampully, 2019; Xiao & Kumar, 2019). Research suggests that technology may be particularly valuable when complementing human strengths, such as empathy, judgment, and creativity, rather than substituting them (Bowen, 2024; Marinova et al., 2016).

The tension between automation and augmentation is central to service theory and practice. Theoretically, there is a need to understand how service agents and technology interact to co-produce value (Davies, Coole, & Smith, 2017; Edvardsson, Tronvoll, & Gruber, 2011; Grönroos, 2011; Trist & Bamforth, 1951; Vargo & Lusch, 2008). Managerially, misreading augmentation as automation can cause an underinvestment in service agent development, suboptimal technology adoption, and missed synergies between human judgment and digital tools (George, Baskar, & Srikaanth, 2024; Montobbio, Staccioli, Virgillito, & Vivarelli, 2023). Conversely, a deeper understanding of augmentation can help service firms address challenges such as task reallocation, labor quality impacts, cognitive skills effects, as well as increased time

pressure and technology overload (Carroll & Conboy, 2020).

Despite the relevance of this rapid technological transformation, the service literature lacks a comprehensive overview of how technology augments service agents. Existing reviews focus on human services, excluding technology (D. D. Walker et al., 2023), technology-driven delivery where automation replaces humans (Chi et al., 2020), isolated technology applications (Shah et al., 2023), or AI effects on organizational level changes like worker roles and collaboration (Bankins, Ocampo, Marrone, Restubog, & Woo, 2023). Furthermore, previous work often addresses specific tools or applications, rather than examining the underlying technological capabilities that transcend tools and sectors for augmenting service agents (M.-H. Huang & Rust, 2018; Kraus et al., 2024; Marinova et al., 2016). The literature calls for a cross-disciplinary understanding that centers service agents in technology-enabled service and clarifies how augmentation occurs across interaction settings and technologies (Larivière et al., 2017; Odekerken-Schröder, Mennens, Steins, & Mahr, 2022).

To address this void, we adopt socio-technical systems theory (Trist & Bamforth, 1951) to interpret the interplay between technological augmentation and service customization. This perspective moves beyond a purely functional classification, towards an understanding of how service agents and technology co-create value. Based on socio-technical systems theory and the literature on service customization and augmentation, we emphasize that technologies are not merely tools for service automation, but rather enablers of context-sensitive, high-quality service experiences. Thus, our review theorizes augmentation in customer service interactions as a dynamic phenomenon distinct from automation and self-service technology.

This shift from automation to augmentation is motivated by theoretical and empirical evidence that technologies add value when complementing and supporting the uniquely human skills of service agents (M.-H. Huang & Rust, 2018; Larivière et al., 2017; Marinova et al., 2016). Our review builds on this perspective by focusing on core technological capabilities underlying service augmentation: AI, as the overarching paradigm; machine learning (ML) and deep learning, for perception and decision support; and natural language processing (NLP), for interpreting and generating human language. These domains are foundational elements of modern AI in service interactions (Hirschberg & Manning, 2015; Jordan & Mitchell, 2015; LeCun et al., 2015; Russell & Norvig, 2016). This highlights how evolving digital capabilities are reshaping service augmentation.

Accordingly, we examine how technology is integrated into frontline service roles,

and how this affects service processes and outcomes across all actors in the technological service triad (Belanche, Belk, Casaló, & Flavián, 2024; Larivière et al., 2017; Odekerken-Schröder et al., 2022). Unlike prior work examining digital transformation broadly or focusing on self-service and automation (Chi et al., 2020; Shah et al., 2023; van Doorn et al., 2016), to our knowledge, this review is the first to systematically synthesize empirical findings on the augmentation of human service agents in service delivery, contributing to the literature in at least four ways.

First, we provide a human-centric lens, positioning the service agent as a key actor in technology-enabled service delivery. This reflects recent calls to examine service agents' perspectives in service systems and explore how technology can complement human service qualities (Larivière et al., 2017; Marinova et al., 2016).

Second, we contribute to the understanding of technology-enabled augmentation of service agents, organizing the literature into core themes mapped onto the stages of the customer interaction process (M.-H. Huang & Rust, 2018; Zapf, Isic, Bechtoldt, & Blau, 2003). Following a thematic mapping approach (Clarke & Braun, 2014), we categorize technologies into six themes: pre-service optimization, interaction intelligence, service agent well-being, service agent monitoring, emotion work, and augmentation. These themes reflect optimization of the targets and purposes, informed by the service agent experience as "the totality of cognitive, emotional, behavioral, sensorial and social responses that result from interactions with other parties (e.g., customers, and technology)" (Larivière et al., 2017, p. 242).

Third, we anchor these themes in a framework derived from socio-technical systems theory (Trist & Bamforth, 1951), mapping them along two dimensions: 1) service customization, as the capability to tailor and adapt service to individual needs and situations (M.-H. Huang & Rust, 2018), and 2) technological augmentation, which describes the extent technology actively supports or amplifies human service work (Larivière et al., 2017; Marinova et al., 2016). This systematic approach shows how technology in human-to-human interactions evolves from manual processes to adaptive, intelligent systems that support service interactions (M.-H. Huang & Rust, 2018). Beyond organizing existing knowledge, this framework identifies critical gaps and tensions, revealing promising research opportunities.

Finally, the multidisciplinary approach integrates insights from service management and marketing, organizational behavior, information systems, and computer science (Bardhan et al., 2010; De Bruyn, Viswanathan, Beh, Brock, & Von Wangenheim, 2022; Kraus et al., 2024). Thereby, the chapter provides a comprehensive

understanding of technology-augmented customer interactions. Building on our conceptual framework's two dimensions, the research agenda identifies future research directions aligned with the reviewed themes. It highlights important questions around human-technology collaboration, service agent well-being, personalization dynamics, and system design.

2.2 Methodology

2.2.1 Search strategy

Figure 2.1 illustrates the article identification process following the PRISMA flowchart. In the first step, a literature search was conducted using Scopus and Web of Science, selected for their extensive coverage of peer-reviewed research across disciplines, ensuring both technical and business-related publications were included (Falagas, Pitsouni, Malietzis, & Pappas, 2007).

Guided by our conceptual anchors, we restricted our review to studies examining technologies that support, extend, or otherwise enhance human service agents, explicitly excluding research on full automation or self-service technologies. Our search string incorporated core technological domains underpinning service augmentation: artificial intelligence (AI), machine learning (ML), deep learning, and natural language processing (NLP) (Hirschberg & Manning, 2015; Jordan & Mitchell, 2015; LeCun et al., 2015; Ledro, Nosella, & Vinelli, 2022; Russell & Norvig, 2016). We also included broader technological search terms such as "technology", "automat*", "intelligent", and "smart" to ensure coverage of the diverse language used in both business and technical literature, while remaining anchored in the goal of understanding augmentation rather than substitution. We deliberately excluded tool-specific terms, such as chatbots or virtual agents, which are often designed for automation and customer self-service (Marinova et al., 2016; Xiao & Kumar, 2019). In contrast, augmentative systems, such as information extraction or decision support tools, often assist service agents and are more accurately identified through their underlying AI capabilities. This approach broadens coverage while maintaining focus on augmentation over substitution.

However, this preliminary analysis also revealed substantial differences in terminology. Service-relevant publications in the beta sciences often use highly technical terms (e.g., "machine learning", "natural language processing") and situate their

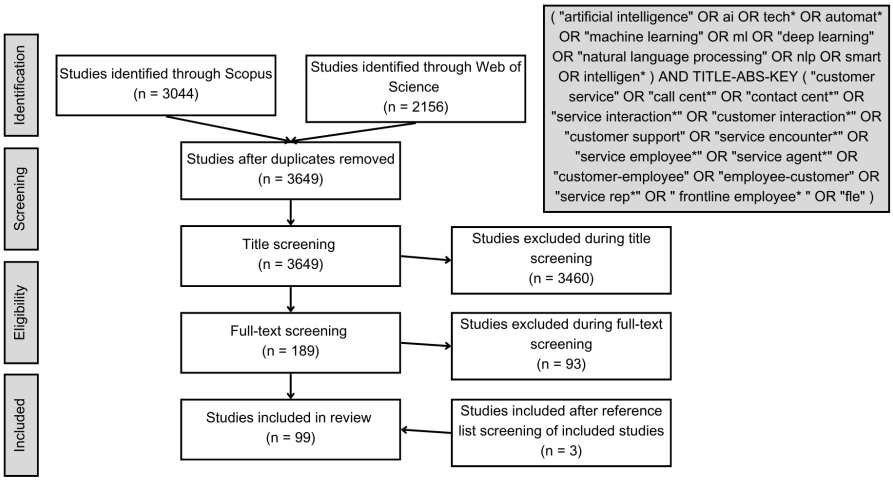

Figure 2.1: Flow diagram of search process

studies in narrowly defined, often technology-centric contexts, mainly “call centers” and “contact centers” (C. Ahmed, ElKorany, & ElSayed, 2023; Borges et al., 2021; Seng & Ang, 2018). These papers rarely reference commonly used terminology in service research and neighboring fields. To capture the service interaction context in these fields, we also needed to include the respective keywords specifically. In contrast, service literature tends to reference broader, conceptually anchored terms, such as “service interaction”, “customer service”, or “frontline employee” to capture a wider range of empirical contexts (A. Fan & Mattila, 2020; Odekerken-Schröder et al., 2022).

Accordingly, our search string is composed of two groups of terms that must co-occur. The first group consists of search terms that capture the technology component of our review objective: (“artificial intelligence” OR “ai” OR “machine learning” OR ml OR “deep learning” OR “natural language processing” OR “nlp” OR “tech*” OR “automat*” OR “smart” OR “intelligen*”). The second group of search terms combines general terminology that detects empirical work on service interactions across both face-to-face and technology-mediated settings and specific service context terms that are prevalent in the beta sciences: (“customer service” OR “service interaction*” OR “customer interaction*” OR “customer support” OR “service encounter*” OR “service employee*” OR “service agent*” OR “customer-employee” OR “employee-customer”

OR “service rep*” OR “frontline employee*” OR “FLE” OR “call cent*” OR “contact cent*”). This approach ensures that our review captures applications of augmentative technology in human service interactions while minimizing the risk of missing relevant studies due to terminological or disciplinary barriers.

Articles required at least one search term from both groups in the title, abstract, or keywords, ensuring both a service context and a technological aspect. Papers published between January 2015 and April 2025 were included in our analysis. The year 2015 was chosen as the starting point, as it marks the time when breakthrough advancements in deep learning and neural networks laid the foundation for today’s transformative AI technologies (Silver et al., 2016). Papers were included only if they examined technologies explicitly designed to augment, support, or enable human service agents, rather than substitute their role in the service process. Further inclusion criteria included peer review status (fully peer-reviewed), language (English), and quality (published in Q1 or Q2 journals, as per the SCImago Scientific Journal Ranking) to ensure high-quality and impactful research (González-Pereira, Guerrero-Bote, & Moya-Anegón, 2010). This process resulted in 5,200 articles, out of which, after removing duplicates, 3,649 papers remained.

In the second step, all titles and abstracts were screened independently by two expert coders. In cases of disagreement, a third coder made the final decision following a conservative approach (i.e., including a paper when in doubt). Papers that did not meet the criteria were excluded, such as those that exclusively featured chatbots or conversational agents, which replaced rather than assisted human service agents. This screening yielded 189 papers, which underwent full-text screening in step 3 to assess eligibility against predefined inclusion and exclusion criteria. Key criteria included a focus on customer service, human-to-human interactions, technology usage, and empirical methodology. This process resulted in 96 included papers, with three additional papers identified through reference list screening, totaling 99.

2.2.2 Bibliometric analysis

First, keyword analysis explored the topical focus of the selected studies. To ensure consistency, we standardized similar keywords (e.g., “NLP” and “natural language processing”). Co-occurrence frequencies were calculated based on the joint appearance of keywords within the same study, forming a weighted network in Python using the networkx library. The network layout was computed using the Fruchter-

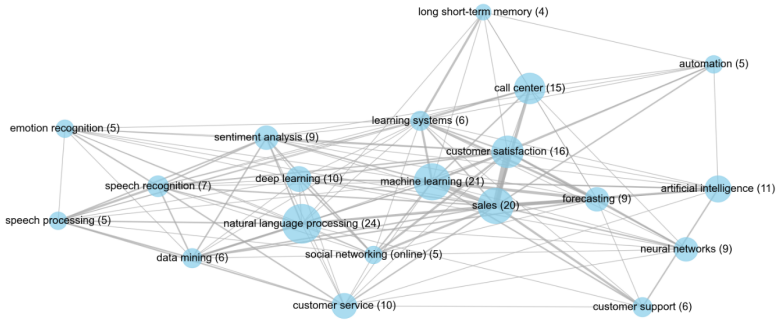

Figure 2.2: Keyword co-occurrence network graph

Note. The dots represent the keywords, with larger dots indicating higher frequencies of the keywords. The lines between the keywords represent their relation, with the keywords co-occurring more frequently for wider lines.

man–Reingold force-directed algorithm (spring_layout), with edge weights representing co-occurrence strength. Nodes with a degree below a threshold (55) were filtered out to highlight the most salient relationships. Figure 2.2 shows the resulting co-occurrence graph, where dots represent keywords and connecting lines show relationships. Dot size corresponds to keyword frequency, with larger dots indicating more occurrences (exact numbers between brackets). Line thickness indicates the co-occurrence frequency, where thicker lines denote keywords that appear more frequently together. As a result, “natural language processing” appeared most frequently (24). Next are “machine learning” (21) and “sales” (20), followed by “customer service” (16) and “call center” (15). Most other keywords are focused on modeling and techniques, including “artificial intelligence”, “neural networks”, and “speech recognition”. The visualization highlights central research domains and illustrates how key concepts interconnect, revealing dominant focus areas and topic clustering within the literature.

We conducted a systematic thematic analysis of the literature on technological interventions in human-to-human customer service, following Braun and Clarke’s (2006) six-phase approach. After familiarizing ourselves with the studies through multiple screenings, we generated initial codes capturing key technologies, implementation approaches, and outcomes. These codes were manually clustered into potential themes, which were iteratively refined through team discussions to ensure internal homogeneity and external heterogeneity.

In building the themes, we focused not only on the type of technology but also on its role within the service context. We examined how technologies interact with both social and technical elements, their location in the service process, who they affected (e.g., the customer or the service agent), and what kind of work they changed or supported. This helped us distinguish, for example, between technologies that assist service agents in understanding conversations and those that monitor their performance, even if both rely on similar underlying methods. As coding progressed, recurring patterns emerged in how technologies shape interactions, tasks, or working conditions, informing themes that reflect different ways technology is embedded in human-to-human service. Each paper was assigned to a single theme to highlight its primary contribution. Although some studies cover multiple aspects, this approach shows key trends in the literature. Assigning just one theme ensures a distinct perspective on technological interventions in customer service. This analysis yielded six themes that represent technology integration in human-to-human customer service, as illustrated in Table 2.1.

The first theme, pre-service optimization, focuses on the initial phase of customer interactions, aiming to minimize waiting times and match customers with the best service agent and channel (Z. Liu et al., 2019; Sandra, Prabowo, Gaol, & Isa, 2024). This involves scheduling service agents based on expected demand and routing customers to optimize the journey before human interaction begins (Ilk, Shang, & Goes, 2020; X. Sun, 2019).

The second theme, interaction intelligence, examines technologies that analyze customer interactions by interpreting conversational data to reveal communication patterns and needs (S. Fan & Ilk, 2020).

The third theme, service agent well-being, focuses on how technology affects service work environments, highlighting both positive and negative impacts on service agent stress and well-being (Henkel, Bromuri, et al., 2020; Y. Huang & Gursoy, 2024).

The fourth theme, service agent monitoring, involves the process of observing and analyzing the behavior and performance of the service agent, including automating feedback and evaluation processes (A. Ahmed, Sivarajah, Irani, Mahroof, & Charles, 2024; Obinna Iheme & Ozan, 2022).

Table 2.1: Overview of the identified themes, descriptions, and related methods

Theme	**Description (Subtopics)**	**Articles**
Theme 1: Pre-service optimization (n = 23)	This theme focuses on the pre-service phase with the goal of minimizing waiting times and ensuring customers are connected to the most suitable service agents or channels. • Predictive resource allocation • Forecasting • Queuing • Service agent or channel matching	Abi Kanaan et al. (2024) Albrecht, Rausch, and Derra (2021) Ali Zaidi, Fraz, Shahzad, and Khan (2021) Bojanić, Delić, and Karpov (2020) Borg, Boldt, Rosander, and Ahlstrand (2021) Bruni, Bianchi, and Papa (2023) Chacón, Koppisetti, Hardage, Choo, and Rad (2023) Ebadi Jalal, Hosseini, and Karlsson (2016) Ilk et al. (2020) Z. Liu et al. (2019) Manno, Rossi, Smriglio, and Cerone (2023) Marín Díaz, Gómez Medina, and Aijón Jiménez (2025) Mohammed (2017) Montgomery, Damian, Bulmer, and Quader (2018) Namli et al. (2021) Legros (2021a) Legros (2021b) Sandra et al. (2024) Schecter, Wowak, Berente, Ye, and Mukherjee (2021) X. Sun (2019) X. Sun and Liu (2023) J. Yang et al. (2024) Yu, Xu, and Tang (2024)

Theme	Description (Subtopics)	Articles
Theme 2: Interaction intelligence (n = 22)	This theme examines the technologies that capture and analyze conversational details from customer interactions. • Speech recognition and transcription • Customer recognition • Customer classification • Topic modeling	Bost, Senay, El-Bèze, and De Mori (2015) Büyük (2024) Cai et al. (2025) S. Fan and Ilk (2020) Galal, Yousef, Zayed, and Medhat (2024) González-Docasal et al. (2020) Q. Han, Yang, Lin, and Qin (2024) Kazanci (2025) H. Lin et al. (2023) X. Ma, Deng, Du, and Pei (2023) Oraby, Bhuiyan, Gundecha, Mahmud, and Akkiraju (2019) Papadia, Pacella, and Giliberti (2022) Papadia et al. (2023) Plaza, Pawlik, and Deniziak (2021) Poczeta, Plaza, Zawadzki, Michno, and Krechowicz (2024) Rajaobelina, Brun, and Ricard (2019) Saberi, Theobald, Hussain, Chang, and Hussain (2018) Shahin, Chen, Hosseinzadeh, Maghanaki, and Eghbalian (2024) Sikveland and Zeitlyn (2017) Valizada, Akhundova, and Rustamov (2021) Vo, Liu, Li, and Xu (2021) Y. Zhou, Fei, Yang, and Kong (2025)

Theme	Description (Subtopics)	Articles
Theme 3: Service agent well-being (n = 10)	This theme explores how technologies are related to the service agents' stress levels and well-being. • Service agent stress • Service agent well-being	Breit, Egeland, Løberg, and Røhnebæk (2020) Bromuri et al. (2021) Choi and Kim (2025) Y. Huang and Gursoy (2024) Oder and Béland (2025) Pacella, Vasco, Papadia, and Giliberti (2024) E. Park, Lee, Han, Diefendorff, and Lee (2024) Phillips et al. (2025) J. Yang et al. (2024) S. Zhou, Yi, Rashiah, Zhao, and Mo (2024)
Theme 4: Service agent monitoring (n = 7)	This theme studies the technologies used to monitor the behavior and performance of service agents. • Service agent assessments • Service agent behavior	A. Ahmed, Toral, Shaalan, and Hifny (2020) A. Ahmed et al. (2021) A. Ahmed et al. (2024) C. Ma and Ye (2022) Obinna Iheme and Ozan (2022) Rees et al. (2021) Shabanpour et al. (2023)

Theme	Description (Subtopics)	Articles
Theme 5: Emotion work (n = 14)	This theme centers around understanding, recognizing, and managing customer sentiments and emotions. • Emotion recognition • Sentiment detection	C. Ahmed et al. (2023) Alam, Danieli, and Riccardi (2018) Al-Mutawa and Al-Aama (2024) Ashtar, Yom-Tov, Rafaeli, and Wirtz (2023) Benayas, Sicilia, and Mora-Cantallops (2024) Carvalho, Oliveira, and Silva (2023) D. Chen, Zhengwei, Jintao, and Khanal (2024) De Cleen, Baecke, and Goedertier (2025) Guo, Li, Liu, and Xu (2024) Labat, Demeester, and Hoste (2024) Badshah et al. (2019) Pérez-Toro, Vásquez-Correa, Bocklet, Nöth, and Orozco-Arroyave (2023) Seng and Ang (2018) Yurtay, Demirci, Tiryaki, and Altun (2024)

Theme	Description (Subtopics)	Articles
Theme 6: Augmentation (n = 23)	This theme highlights the use of technology in collaboration with the service agent or as an augmentation tool to enhance the interaction. • Emotion recognition • External devices (Apps, robots, glasses) • Information structures	Arwin, Halldórsson, and Hellström (2024) M. Blaurock, Büttgen, and Schepers (2024) Brynjolfsson, Li, and Raymond (2025) Choi (2018) De Gauquier, Willems, Cao, Vanderborght, and Brengman (2023) Dolata, Agotai, Schubiger, and Schwabe (2020) A. Fan and Mattila (2020) Gnewuch, Morana, Hinz, Kellner, and Maedche (2023) Henkel, Bromuri, et al. (2020) L. L. Huang, Chen, and Chan (2024) J. J. Kim et al. (2025) Le, Sajtos, Kunz, and Fernandez (2024) Leiño Calleja, Schepers, and Nijssen (2025) Levi-Bliech, Pliskin, and Fink (2020) X. Lin, Wang, Shao, and Taylor (2024) H.-F. Lin (2025) Moliner-Tena, Callarisa-Fiol, Sánchez-García, and Rodríguez-Artola (2024) Poots, Morgan, Woolf, and Curcuruto (2024) Pöyry, Holopainen, Parvinen, Mattila, and Tuunanen (2024) Sheng, Natalia, and Rusfian (2024) Wei, Lu, Cheng, Jiang, and Liu (2022) L. Wu, Fan, and Mattila (2015)

The fifth theme, emotion work, shows the role of technology in recognizing and managing customer emotions (C. Ahmed et al., 2023; Guo et al., 2024).

Finally, the sixth theme, augmentation, highlights how technology can act as a partner during service, augmenting the human service agent. These tools aim to optimize interactions between customers and service agents by combining human expertise with technological support, thereby improving efficiency (Dolata et al., 2020).

These six themes demonstrate how technology is currently integrated into human-to-human customer service, as illustrated in Table 2.1. The findings within each category are explored in the results section.

Based on these themes, we can explore further insights. First, we analyze both yearly distribution per theme and total annual publications (see Figure 2.3). Overall, there has been a steady increase in publications on these topics, with a slight dip observed in 2022. As articles are only included up to April 2025, a growth trend is expected, as denoted by the star in Figure 2.3. Notably, themes 6 (augmentation) and 3 (service agent well-being) peak in 2024, highlighting both the recency and rising importance of these topics. This trend reflects growing research interest not only in applying technical methods to service data but also in how technology supports augmentation and service agent well-being.

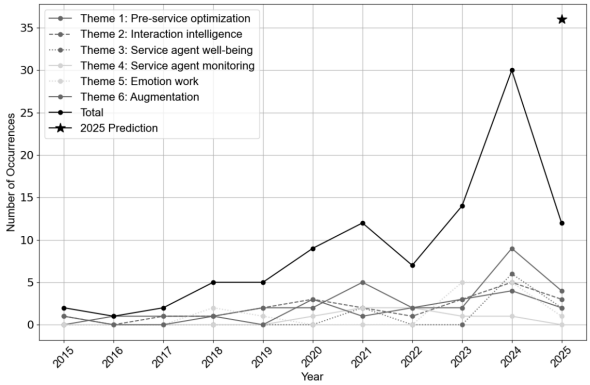

Figure 2.3: Distribution of papers over time: total and by theme
Note. The star at 2025 indicates a projected total of 36 occurrences, estimated based on the count up to April (12 occurrences) and the upward trend observed in previous years, particularly the sharp increase in 2024.

Subsequently, we examined the types of modalities employed in the articles (see Figure 2.4). The majority rely on text-based communication, such as live chat or tran-

scribed audio. This reflects the continued centrality of written communication in organizational service interactions. Speech is the second most commonly used modality, typically analyzed in phone-based customer service. While speech data enables the analysis of tone and paralinguistic cues, it is often treated separately from textual data, limiting opportunities to understand how service agents shift across modalities in real-time or utilize them in tandem. Only six studies combine text and speech, suggesting that multimodal analysis remains relatively rare, despite its potential to capture the complexity of frontline service interactions. This scarcity may stem from methodological difficulties or data availability constraints, but it also signals a missed opportunity to better reflect the hybrid nature of modern service delivery. Remarkably, just one study includes video as a modality. This underrepresentation may reflect both the limited availability in practice and the methodological or privacy-related constraints of studying such data. It also indicates that, to date, service research has focused predominantly on text and speech, with minimal attention to visual modalities.

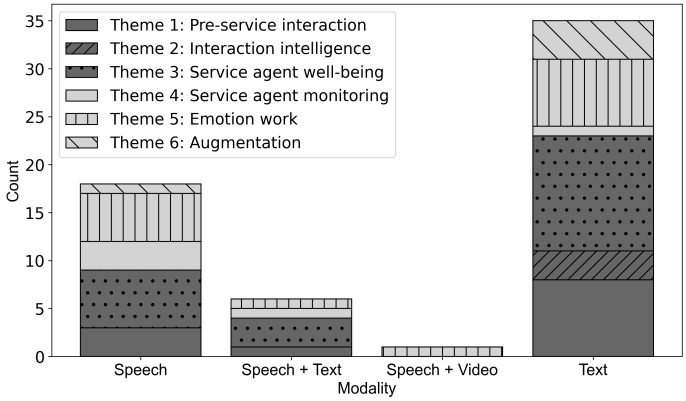

Figure 2.4: Distribution of modalities: total and by theme

Finally, we examined model types (see Figure 2.5). Most studies rely on ML models, particularly neural networks. ML is a type of AI where models learn patterns from data to make predictions about unseen data (Janiesch et al., 2021). Neural networks are ML models inspired by how the human brain processes information, breaking information into smaller parts processed in multiple steps to identify patterns (Janiesch et al., 2021). Additionally, studies explore various model types, including robots, mobile applications, augmented reality systems, and simulation-based studies, high-

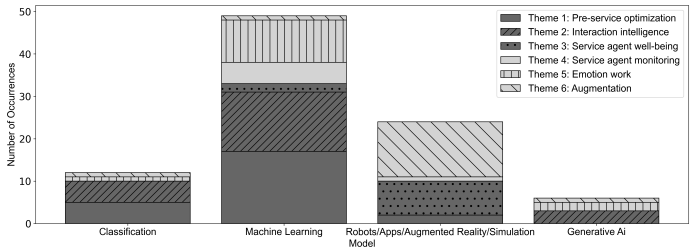

Figure 2.5: Distribution of types of models: total and by theme

lighting the diverse applications of technology in this field. Then, classification models, including statistical approaches like regression, are the third most common type. Notably, few studies utilize a generative AI model, indicating this area remains in an early exploration stage. Except for one study in 2023, all generative research was published later, highlighting its recent emergence.

2.3 Results

This results section explores the six themes identified from our systematic review: pre-service optimization, interaction intelligence, service agent well-being, service agent monitoring, emotion work, and augmentation. Here, we highlight the distinct contributions of each theme to understanding the broader landscape of service augmentation, as well as their interdependencies and boundary conditions. Through this exploration, the themes will serve as a structured framework for analyzing the findings and their implications.

2.3.1 Theme 1: pre-service optimization

Pre-service optimization uses technology to route customers to the most suitable service agent, aiming to reduce wait times and churn before the actual interaction begins (Haenlein & Kaplan, 2012; Vo et al., 2021). Service agent scheduling is crucial, as insufficient staffing increases wait times. These allocations rely on call arrival forecasting (Albrecht et al., 2021; Chacón et al., 2023; Ebadi Jalal et al., 2016; Manno et al., 2023). Advanced approaches include predicting abnormal call volumes using service agent skill-based methods (Mohammed, 2017) and proactively detecting issues to prevent customer calls (Namli et al., 2021). To handle peak periods, service

agents may be scheduled on demand (X. Sun & Liu, 2023).

Alongside fully human service agent scheduling, the combination with virtual agents presents a distinct optimization challenge. While customers prefer human interaction, virtual agents offer unlimited availability (Yu et al., 2024). Research suggests that scheduling strategies anticipating demand yield better outcomes than reactive approaches (Legros, 2021b). Furthermore, some customers or problems require a higher level of urgency and should therefore be prioritized. This urgency can be determined through various measures, such as customer emotions, message content, or customer priority (Abi Kanaan et al., 2024; Bojanić et al., 2020; Marín Díaz et al., 2025; J. Yang et al., 2024).

Traditional routing through first-line service agents to specialists has revealed inherent biases, as familiar service agents receive disproportionately high call volumes, creating inertia effects that impact performance (Schecter et al., 2021). Research shows that automated routing based on text analytics, service agents' skills, or customers' personality traits improves resolution rates and customer satisfaction (Ilk et al., 2020; Sandra et al., 2024; X. Sun, 2019).

Research on service center operations emphasizes balancing between inbound and outbound communications, while also optimizing service agent idle time (Legros, 2021a). Machine learning supports this through cross-channel routing optimization (Z. Liu et al., 2019). For asynchronous channels, such as emails and support tickets, studies have shown the value of urgency ranking, topic categorization, and automated resolution suggestions, linking this first theme to the second theme (Ali Zaidi et al., 2021; Borg et al., 2021; Bruni et al., 2023; Montgomery et al., 2018).

2.3.2 Theme 2: interaction intelligence

Customer service conversations contain valuable insights that can be utilized both during and after interactions. For phone-based interactions, transcribing speech facilitates information acquisition, as transcriptions can be easily read and analyzed. Additional steps, such as pre- and post-processing, can enhance transcription accuracy (Galal et al., 2024; X. Ma et al., 2023; Plaza et al., 2021; Valizada et al., 2021; Y. Zhou et al., 2025).

After transcription, NLP techniques enable more detailed analysis. Topic modeling categorizes conversations into one or more predefined categories, identifying common customer questions and issues (Bost et al., 2015). Based on these topics,

transcripts can be automatically classified into cohesive and well-separated groups, suggesting similar behavior or follow-up steps within each group (Papadia et al., 2022, 2023). Similarly, intent classification can detect the underlying goals of the customer, which can be extracted from interactive voice responses or from the call itself (Cai et al., 2025; Kazanci, 2025). Furthermore, conversations can be summarized, allowing faster and more compact information usage (Büyük, 2024; Q. Han et al., 2024; H. Lin et al., 2023).

Additionally, speaker information, including personal details and customer problems, can be extracted (Hathaway, Emadi, & Deshpande, 2021; Saberi et al., 2018). Once identified, customers can be classified into different types, allowing for a more personalized approach (Marín Díaz et al., 2025; Rajaobelina et al., 2019). Similarly, detailed information extraction can enhance customer understanding (S. Fan & Ilk, 2020; Oraby et al., 2019). Audio features can help detect tension or conflict, where a customer might be upset, frustrated, or trying to assert themselves (Sikveland & Zeitlyn, 2017). More generally, frameworks can combine methods to enhance analysis information (S. Fan & Ilk, 2020; González-Docasal et al., 2020). As customer language and behavior evolve, regular retraining of AI models further enhances classification accuracy (Poczeta et al., 2024).

In summary, interaction intelligence proposes technologies to analyze and interpret service interactions (Hathaway et al., 2021). By extracting detailed patterns, it enhances the understanding of communication dynamics and improves the effectiveness of customer service and service agent performance, linking to other themes such as service agent well-being (Theme 3) and service agent monitoring (Theme 4). Moreover, the analytical capacity developed within interaction intelligence can directly support Theme 1 by informing early-stage service processes with information extracted from previous customer interactions.

2.3.3 Theme 3: service agent well-being

Service agents form the backbone of customer service (Marr, 2024). Ensuring their satisfaction to retain them remains a significant challenge. Turnover rates are high (Zito et al., 2018), and training new service agents is both time-consuming and costly (Hillmer, Hillmer, & McRoberts, 2004). Here, technology has been suggested to support service agents (Bankins et al., 2023).

One way to support the service agent is by improving their workload and well-

being (Choi & Kim, 2025; Pacella et al., 2024). Here, stress levels play a crucial role in shaping both performance and overall well-being, making it essential to identify and effectively address stress (De Ruyter, Wetzels, & Feinberg, 2001). In one study, a machine learning model predicted the service agent's stress in real-time based on the customer's emotion patterns (Bromuri et al., 2021). ML can evaluate the emotional workload of service agents, providing managers with insight into stress levels and enabling timely interventions (E. Park et al., 2024). Additionally, technology can improve well-being directly. Service robots were proposed to allow service agents to adjust their physical and purpose-related work through task allocation strategies, which helps restore both physical and work well-being dimensions (Phillips et al., 2025).

However, technology can also introduce new or additional stressors. For example, greater system availability may shift responsibilities to digital platforms, pressuring service agents to remain constantly connected (Breit et al., 2020). While increasing transparency, this necessitates more cautious communication, as statements could be used against them or the organization. AI awareness has also been shown to cause negative emotional responses, including counterproductive work behavior (S. Zhou et al., 2024).

In contrast, anthropomorphism of service robots can increase service agent awareness of smart technologies, helping them work with these tools (Y. Yang, Chi, Bi, & Xu, 2024). However, it may also reduce emotional warmth, weakening service agents' positive connection to the technology and potentially affecting motivation and well-being. Another study examines the relationship between AI integration and service agents' proactive service behaviors, yielding mixed results (L. L. Huang et al., 2024). While perceiving AI integration as a positive challenge enhances proactive service behaviors, it also increases job insecurity, which in turn decreases these behaviors.

Similarly, research highlights how generative AI tools increase emotional labor among low-skilled workers while also limiting their ability to contextualize these challenges within broader labor market dynamics (Oder & Béland, 2025). This highlights the importance of supporting service agents during technology implementation through targeted measures, such as educational campaigns and upskilling programs. Without suitable support, these technologies can increase stress and ineffective coping strategies. Moreover, it can contribute to performance issues, and addressing these issues is crucial for ensuring consistent, high-quality customer service and pre-

venting negative interactions (De Ruyter et al., 2001).

2.3.4 Theme 4: service agent monitoring

Service agents have highly structured jobs, balancing multiple responsibilities (Tovar, 2021; Zapf et al., 2003), bridging operations, sales, marketing, and technology (Choi, 2018). To support this multifaceted role, organizations increasingly use technological solutions that monitor, visualize, and enhance service agent performance.

Service agents' behaviors can be visualized to highlight the strengths, weaknesses, trends, and training opportunities (Rees et al., 2021). Subsequently, negative customer interactions can be reduced by understanding the behavior of service agents. For example, service agent malpractice and service sabotage can be detected automatically, allowing for early interventions (C. Ma & Ye, 2022; Obinna Iheme & Ozan, 2022). This underscores the need for a supportive work environment, where negative behavior is addressed early but in a supportive way. Given the issues with manual feedback, such as limited and delayed feedback, technologies have been introduced into this process (A. Ahmed et al., 2021). These methods have been applied to classify service agent productivity and identify subjective calls, resulting in more equitable processes based on predefined criteria (A. Ahmed et al., 2021, 2024, 2020; Shabanpour et al., 2023).

This fourth theme closely relates to Theme 2, as both leverage data-driven analysis to improve interaction quality. While interaction intelligence analyzes conversational content to understand customer behaviors, service agent monitoring evaluates the behavior and performance of service agents. Integrating these themes enables a holistic view of how service agent actions align with customer expectations. Insights from this theme contextualize service agent behavior, supporting targeted coaching and continuous improvement. This synergy enhances service effectiveness and customer satisfaction.

Another critical relationship exists between Themes 3 and 4. Monitoring service agents can enhance their well-being by receiving timely, personalized feedback and participating in targeted interventions, thereby reducing stress and improving job satisfaction (E. Park et al., 2024). However, pervasive surveillance may undermine service agents' sense of agency and can lead to anxiety and decreased motivation, resulting from constant observation and evaluation (C. Ma & Ye, 2022). Therefore, studying these themes together is crucial for understanding how monitoring practices

affect both psychological health and performance.

2.3.5 Theme 5: emotion work

Customer emotions have a significant influence on service experiences, affecting both positive and negative service outcomes (Mattila & Enz, 2002). Therefore, service agents' interpersonal emotion regulation skills are crucial in shaping customer interactions (Reeck & Onuklu, 2022; Zaki & Williams, 2013). However, effectively managing these emotions requires recognizing them first. Here, two approaches are reflected in the literature: sentiment detection and emotion recognition.

First, customer sentiment detection provides insights into customers' affective states (Carvalho et al., 2023). Recent studies have utilized sentiment analysis and customer satisfaction recognition to gain deeper insight into customer interactions (C. Ahmed et al., 2023; Al-Mutawa & Al-Aama, 2024; Ashtar et al., 2023; Benayas et al., 2024; Carvalho et al., 2023; Yurtay et al., 2024). A large-scale study of call center interactions revealed that positive customer sentiment has a significant impact on both satisfaction and intention to recommend. In contrast, negative emotions have a greater effect on recommendations than on satisfaction. Service agent sentiment has a lower influence, and emotional matching between customers and service agents is generally beneficial (De Cleen et al., 2025).

Second, emotion recognition studies adopt a more granular approach, using diverse data modalities and methods. For instance, dialogues can be annotated with emotions, valence-arousal-dominance scores, and service agent response strategies (Alam et al., 2018; Labat et al., 2024; Pérez-Toro et al., 2023). Expanding the types of multimodal analysis (Badshah et al., 2019; Guo et al., 2024), a novel approach was introduced by combining video and audio (Seng & Ang, 2018). Taking it one step further, machine learning can be utilized to analyze customer emotion patterns and predict service agent stress in real-time (Bromuri et al., 2021). Moreover, automatic emotion recognition can aid service agents, enhancing their effectiveness in regulating emotions during interactions (Henkel, Bromuri, et al., 2020). This demonstrates the progress in emotion recognition, underlining their potential to support service agents in their roles.

This fifth theme complements both Theme 1, pre-service optimization, and Theme 2, interaction intelligence, by adding an emotional dimension. While Theme 1 focuses on pre-call strategies, and Theme 2 analyzes the content and patterns during

the call, Theme 5 enriches information by capturing how customers express themselves emotionally. This emotional information can then be linked back to Theme 1, allowing for the anticipation of potential emotional states and optimizing wait times and routing accordingly. Together, these themes offer a comprehensive view of customer experience, from anticipating needs to managing emotional dynamics during interactions.

Theme 5 also connects with Theme 3, as both address the emotional dimensions of interactions. Emotion recognition technologies enable the real-time detection of customer affect, equipping service agents to tailor responses and manage challenging interactions. This emotional insight can reduce uncertainty and support emotional regulation, thereby enhancing emotional resilience and job satisfaction. However, they may also heighten emotional labor, as service agents are expected to continually respond to emotional cues, which can potentially lead to increased fatigue and burnout.

2.3.6 Theme 6: augmentation

One of the current key challenges is determining how service agents and customers interact with technology. Understanding the human-AI collaboration is crucial for enhancing all aspects of customer service (Le et al., 2024).

Technology can complement human service agents in various service settings. For instance, in retail, service agent-robot teams have shown that increased attention from robots does not necessarily lead to higher sales (De Gauquier et al., 2023). The effectiveness of such collaboration depends on service agent characteristics and motivation; individuals with positive attitudes and lower anxiety are more likely to collaborate successfully with robots (H.-F. Lin, 2025). Customer perceptions also play a role; frontline robots with a high level of automated social presence enhance impressions of service agent competence, warmth, and overall teamwork quality (Leiño Calleja et al., 2025).

Similar dynamics appear in human-chatbot collaboration. Customers respond more favorably when human involvement is disclosed, often shifting to a more human-oriented communication style (Gnewuch et al., 2023; Wei et al., 2022). However, this shifts more conversations to human service agents, thereby limiting chatbot scalability and raising service agent workload. Customers tend to prefer augmentation settings, where the human remains central to the decision-making process (Le et al.,

2024). Additionally, artificial agents are evaluated more positively when paired with creative human service agents (L. L. Huang et al., 2024).

In a restaurant context, innovative technologies are reshaping the roles of both service agents and customers, with service agents demonstrating a greater awareness of these changes. Role stress impacts psychological empowerment, while collaborative value co-creation between service agents and customers contributes to positive behavioral intentions (J. J. Kim et al., 2025). Research also highlights how collaborative intelligence systems can enhance service outcomes, work meaning, and system adherence, particularly among service agents with limited experience with AI (M. Blaurock et al., 2024).

Information technology can also support service agents (Choi, 2018). One example is digital face-to-face interactions, which can advance relational primary healthcare (Arwin et al., 2024). In telephone triage, digital systems support assessment and decision-making processes, but also require ongoing adaptation to ensure safe and effective outcomes (Poots et al., 2024). Another study investigated Google glasses by service agents at hotel check-in, showing gender as a moderating variable (L. Wu et al., 2015). Other studies report on mixed-reality systems that support natural behaviors, such as writing and handling paper, and integrate them into computer-generated visualizations to enhance transparency (Dolata et al., 2020). Similarly, sales support applications can complement face-to-face interactions, facilitating electronic commerce and leading to more customer visits and purchases (Levi-Bliech et al., 2020).

In addition, conversational agents can be implemented to support the service agent, boosting service agents productivity and English fluency, especially for moderately rare problems, international service agents, and lower-skilled workers (Brynjolfsson et al., 2025). They can also enhance overall work performance and work experience, leading to more polite customer behavior and a reduction in requests to speak to a manager (X. Lin et al., 2024). These augmentation chatbots can also enhance brand engagement and increase usage intention compared to substitution chatbots (Sheng et al., 2024). Moreover, emotion recognition techniques can enable service agents to reflect customer emotions in real-time, facilitating effective interpersonal emotion regulation strategies (Henkel, Bromuri, et al., 2020).

Service robots represent another domain where technology can augment the work of service agents. In hotel settings, collaborative robots are often perceived by customers as complementary and contribute significantly to service outcomes (Moliner-

Tena et al., 2024). Finally, virtual reality can enhance collaboration and problem-solving in service interactions, particularly in knowledge-intensive contexts involving novice decision-makers. It enables more guided and focused interactions between customers and service agents (Pöyry et al., 2024).

This final theme of augmentation acts as a central connector across all other themes, where technologies from each theme can be employed individually or in combination to enhance service agents through technology. Each technology brings various benefits and challenges, influencing the experience of both the customer and the service agent. Despite this diversity, the overarching goal of augmentation is to empower human service agents by complementing their expertise and judgment with technological assistance, ultimately improving the quality, adaptability, and effectiveness of customer service interactions.

In summary, these six themes collectively map the emerging landscape of technological augmentation in human-to-human service interactions. While each theme addresses distinct aspects, the boundaries between them are blurred in practice. For example, technologies enabling emotion recognition (Theme 5) are closely intertwined with efforts to support service agent well-being (Theme 3) and effective monitoring (Theme 4). Likewise, augmentation (Theme 6) serves as an enabler, shaping and being shaped by other themes. Together, these interconnected themes underscore the necessity of integrated socio-technical approaches for achieving both operational excellence and positive human outcomes. Building on these thematic insights, the following section develops a conceptual framework that synthesizes research domains along the dual dimensions of technological augmentation and service customization, providing a foundation for deeper theoretical integration and future research.

2.4 Conceptual framework

To synthesize our findings, we developed a conceptual framework (see Figure 2.6). Our inductive thematic analysis initially revealed two core dimensions: technological augmentation and service customization, which were subsequently anchored within socio-technical systems theory and the literature on service customization and technological augmentation (Marinova et al., 2016; Rust & Huang, 2014; Trist & Bamforth, 1951). This theoretical lens emphasizes that technologies are more than just

tools for service automation; they act as enablers of context-sensitive, high-quality service experiences. Socio-technical systems theory emphasizes the interdependence of social and technical subsystems within organizations, whereby value is created not by technology or human actors alone, but through their ongoing interaction (Trist & Bamforth, 1951). Translated to a service context, technology is embedded within dynamic networks of human service agents, processes, and customer interactions (Larivière et al., 2017; Marinova et al., 2016). This view aligns with service-dominant logic, where value is co-created through the integration of human and technological capabilities (Vargo & Lusch, 2008). Service agents are not passive recipients of technology; they are actors who actively configure and leverage these tools to improve service outcomes.

A core element is technology's role in augmenting service agent capabilities, including judgment, empathy, and complex problem-solving (Bowen, 2024; M.-H. Huang & Rust, 2018; Marinova et al., 2016; Xiao & Kumar, 2019). A second essential element is customization. Service theory and practice converge around a shift from standardized, one-size-fits-all to adaptive and customized experiences to meet the unique needs of both customers and service agents (Payne & Frow, 2005; Pine, Victor, & Boynton, 1993; Rust & Huang, 2014).

Building on these core elements, our framework is structured along two dimensions: (1) technological augmentation (horizontal axis), which ranges from human-led to technology-augmented, denoting the extent to which technology supports, complements, or substitutes for human agency in service interactions (Bowen, 2016; Marinova et al., 2016), and (2) service customization (vertical axis), which ranges from standardized to highly customized and is defined as the extent to which service processes are adapted to individual customer or service agents needs (Payne & Frow, 2005; Pine et al., 1993).

Placing the axes together yields four configurations that capture current practice and likely evolution. The framework depicted in Figure 6 serves as an analytic lens rather than a rigid classification. Each theme can appear in more than one quadrant, depending on its implementation, and can shift as systems and practices evolve. Their positions in the illustration reflect where they most frequently appeared in the reviewed studies: pre-service optimization, interaction intelligence, and service agent monitoring emerged in low-context, high-efficiency applications; service agent well-being and emotion recognition clustered in more human-led constellations where contextual sensitivity dominated; and augmentation was most often discussed in set-

tings that combine technological and human strengths. These placements should not be interpreted as fixed boundaries, but rather as indicative patterns that highlight how the functional roles of technology have been treated in the literature to date.

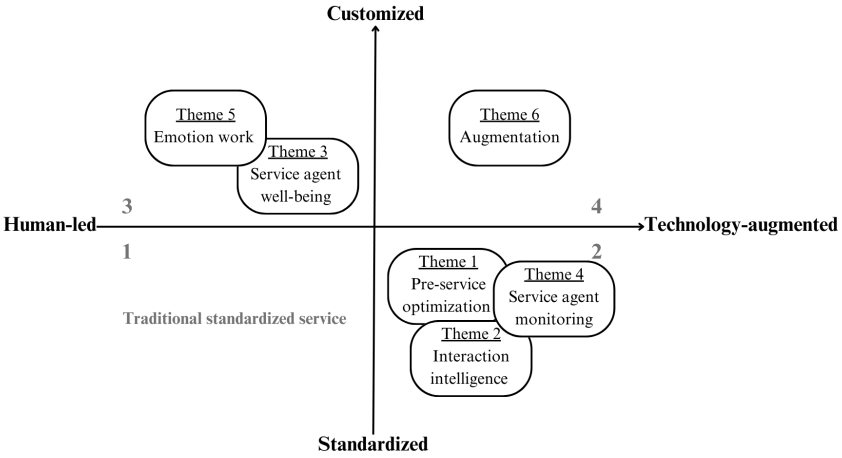

Figure 2.6: Conceptual framework

Note. This framework maps the six core themes identified in the review across two dimensions of service interaction: (1) the degree of technological augmentation (horizontal axis), which ranges from human-led to technology-augmented, and (2) the level of service customization (vertical axis), which ranges from standardized to customized. The framework serves as an analytic lens rather than a rigid classification, with themes placed where they most frequently appear in the reviewed studies. Themes can also span quadrants as implementations and practices evolve.

In line with socio-technical systems theory, technology is an enabler extending human capacity and facilitating real-time customization (Barrett, Oborn, & Orlikowski, 2016; Trist & Bamforth, 1951), This perspective highlights that service augmentation is most effective when technological features fit both functional and emotional demands of frontline service work, enabling service agents to manage information better, respond empathetically, and resolve service failures effectively (Goodhue & Thompson, 1995; Larivière et al., 2017).

While service customization and technological augmentation are often studied independently, socio-technical systems theory emphasizes that integrating these dimensions through aligning the human and technical subsystems is essential for op-

timizing service outcomes (Trist & Bamforth, 1951). Our framework makes this integration explicit, clarifying how different combinations of augmentation and customization facilitate service interactions. Importantly, this alignment also facilitates value co-creation in service, such that the mutually reinforcing nature of technology and human service agents ideally positions the firm to co-create value with customers (Edvardsson et al., 2011; Vargo & Lusch, 2008).

2.4.1 Quadrant 1: standardized & human-led

Quadrant 1 involves human-led service delivered through standardized processes. Service agents execute routine tasks with minimal technical support and low customization, prioritizing efficiency over individual needs. Examples of these interactions can be found in fast-food restaurants, call centers handling simple inquiries, and routine hotel check-ins. Service agents follow well-defined workflows, which reduces training needs and ensures consistent quality standards. However, the lack of customization may limit customer satisfaction, especially when individual preferences or exceptions are not accommodated.

Drawing on socio-technical systems theory, this quadrant represents a traditional service configuration where the role of technology is limited to supporting back-office processes (e.g., scheduling, basic record-keeping), while human service agents operate according to standard scripts (Goodhue & Thompson, 1995; Trist & Bamforth, 1951). Service interactions in this quadrant exemplify the traditional paradigm of mass service, where uniformity, predictability, and scalability are prioritized over flexibility (Bowen, 2024; Pine et al., 1993). Despite the advantages of standardized services, such as operational efficiency, cost control, and process reliability, predefined protocols can be inflexible, limiting opportunities for firms to enhance customer engagement, co-creation, or personalized service interactions.

2.4.2 Quadrant 2: standardized & technology-augmented

Quadrant 2 represents service settings where technology plays a supporting role but lacks adaptability or personalization. Technology is used to manage static, repetitive, or standardized processes, automating routine tasks. However, this technology

is not context- nor customer-specific, as it has limited capacity for dynamic problem-solving due to being designed for consistency and scale.

This quadrant reflects the integration of socio-technical systems theory with service operations literature (Marinova et al., 2016; Trist & Bamforth, 1951), where the technical subsystem is optimized to deliver uniform outputs and ensure reliability. As technology handles more of the routine aspects, it allows humans to focus on personal attention and expertise, thus serving as a supportive tool (Goodhue & Thompson, 1995; Rust & Huang, 2014).

This quadrant includes pre-service optimization, interaction intelligence, and service agent monitoring. Pre-service optimization minimizes wait times, schedules service agents, and connects customers to suitable service agents and channels. Tasks are typically implemented statically using historical data. This approach is often context-general, rather than based on different contexts, such as seasonal differences (Mohammed, 2017) or customer characteristics and needs (Shahin et al., 2024).

Interaction intelligence currently aligns with this quadrant by supporting service environments using generalized, non-personalized models. Systems like transcription software and summary extraction are static models that operate independently of customer context, such as accents, preferences, or emotional tone. They produce broadly applicable outputs but operate independently of human service agents, with results used for post-conversation analysis rather than real-time support. They contribute to efficiency and consistency, not personalization.

Finally, performance monitoring technologies are primarily designed for standardization, compliance, and efficiency. They are used in high-volume service environments to evaluate KPI performance and maintain consistent quality. They align with this quadrant's characteristics, where technology leads in delivering standardized services.

2.4.3 Quadrant 3: customized & human-led

Quadrant 3 incorporates service interactions where humans demonstrate a high degree of customization and adaptability, using limited technological involvement. Service agents rely on recognizing emotional, cultural, and contextual cues to create personalized experiences, while technology serves a supportive, background role (e.g., as a database or information resource) (Prentice et al., 2019; Presbitero, 2016).

Drawing on socio-technical systems theory (Barrett et al., 2016; Trist & Bamforth, 1951) and service customization literature (Payne & Frow, 2005; Pine et al., 1993), this quadrant highlights contexts where human judgment, empathy, and adaptive problem-solving are central. Here, technology primarily acts as an enabler rather than a means of augmentation, reinforcing the social and relational aspects of service. This approach maximizes customer outcomes by enabling flexible, relationship-driven services tailored to unique needs. It supports complex problem-solving and fosters deep human connection. However, it is labor-intensive, difficult to scale, and may lead to inconsistencies in service quality and higher operational costs.

The two themes categorized in this quadrant are service agent well-being and emotion work. For service agent well-being, these technologies help supervisors gain insights into service agents' stress, as various emotional, cognitive, and technological tasks can impact well-being (Pacella et al., 2024). Hereby, supervisors can provide support and enhance service agents' well-being. Second, service agents' interpersonal and emotional intelligence skills are utilized to recognize customer emotions and adjust communication strategies accordingly to de-escalate negative conversations, build rapport, or enhance customer satisfaction. Here, intelligent technologies can recognize customer affect, assisting in emotion regulation.

2.4.4 Quadrant 4: customized & technology-augmented

Quadrant 4 represents advanced service settings where technological systems augment human service agents to deliver personalized, adaptive customer support. Here, digital tools (e.g., AI-driven recommendation systems, real-time analytics, and conversational agents) are deeply integrated into the service process, enabling dynamic customization at scale. Both the technology and the human service agent contribute interactively to co-create value for the customer (Dolata et al., 2020; L. Wu et al., 2015).

From a socio-technical systems theory perspective (Barrett et al., 2016; Trist & Bamforth, 1951) and the literature on technological augmentation (Marinova et al., 2016; Rust & Huang, 2014), this quadrant illustrates the ideal alignment of technical and social subsystems. Such integration maximizes both operational efficiency and the customization of the customer experience, enabling service agents to meet diverse customer needs flexibly and responsively (Edvardsson et al., 2011; Vargo

& Lusch, 2008). The strength of this augmentation lies in delivering high-quality, adaptive services at scale. It may enhance service agent efficiency and effectiveness through intelligent support. However, it requires significant investment in digital infrastructure and may lead to service agent deskilling, overreliance on technology, and increased cognitive load, potentially negatively impacting the service agent. Additionally, it raises concerns around authenticity, trust, and customer privacy.

Theme 6, augmentation, is central to this fourth quadrant, where advanced digital tools actively enhance the ability of service agents to deliver personalized, context-sensitive support. Current research shows early examples including robots in front-line roles, chatbots complementing human service agents, and information technology providing decision support. These technologies augment rather than replace human labor, extending service agent capabilities during interactions.

2.5 Research agenda

Building on our conceptual framework, our research agenda is organized around two dimensions: augmentation and service customization. This approach moves from human-led to technology-augmented services and from standardized to customized delivery, respectively. In the following sections, we elaborate on how progress in each dimension can enable organizations to deliver more personalized, efficient, and responsible service experiences.

2.5.1 From human-led to technology-augmented

The transition from human-led to technology-augmented service represents a paradigm shift toward hybrid environments, where humans and intelligent systems collaborate in real-time. This subsection of the research agenda critically examines this transformation, articulating key avenues for future research that address the evolving roles of technology in customer service interactions. To guide meaningful integration, future research should more closely examine how technologies can be developed to amplify uniquely human strengths and how service agents can acquire the necessary competencies, trust, and fluency to work with these tools.

A relatively new concept in the literature is that of cyborgs (Nyberg, 2009), where human service agents are enhanced by embedding technology directly into their bodies or cognitive processes. Positioned at the far end of this human-led to technology-

augmented continuum, cyborgs merge human and machine capabilities through wearable devices, neural interfaces, or other bio-integrated technologies (Garry & Harwood, 2019; Grewal, Kroschke, Mende, Roggeveen, & Scott, 2022; Nyberg, 2009). This fusion transforms traditional service roles into hybrid entities, enabling deeper collaboration between human and machine intelligence (Theme 6: augmentation). Key research challenges include identifying feasible cyborg technologies for near-term adoption and assessing their ethical and psychological impacts. Cyborg service agents may face challenges related to autonomy, stress, shifting professional identities, and biological risks associated with neural implants (Garry & Harwood, 2019; Kies et al., 2025). Developing ethical frameworks to protect human service agents will be essential (Theme 3: service agent well-being).

Advancing technology-augmented customer services involves overcoming key challenges in building intelligent, adaptive systems. This includes developing multimodal, real-time technologies to create a coherent understanding of interactions. These systems support service agents by enhancing emotional attunement (Theme 5), monitoring well-being (Theme 3) and performance (Theme 4), and enabling responsive, context-aware collaboration. Generative AI further contributes by producing personalized content in real-time (Theme 6), although its integration raises new research questions. Addressing issues such as data fusion, contextual interpretation, and deployment in immersive environments, including robotics and augmented reality, is essential for truly adaptive, human-centered services (Wirtz et al., 2018). Crucially, explainable AI is necessary to support decision-making with transparent and interpretable outputs. Few papers in our review substantively addressed transparency, explainability, or biases, highlighting a critical literature gap. Equally important is embedding these tools into service agents' workflows in a way that preserves user agency and avoids cognitive overload. Future research should investigate how service agents adapt their routines and decision-making processes to increasingly sophisticated technologies.

2.5.2 From standardized to customized

A central focus of our research agenda is the shift from standardized to customized, human-centric service tailored to individual needs and preferences. This subsection explores this transformation and outlines key research directions on technology's evolving role in customer service interactions. A central trajectory in the shift

from standardized to personalized service is increasing use of data-driven personalization strategies. These approaches tailor interactions to individual customers' and service agents' needs, enhancing relevance, satisfaction, and engagement (Theme 2: interaction intelligence) (Ameen et al., 2011). Predictive personalization, which anticipates customer needs before expression, offers particular promise; however, integrating behavioral, contextual, and historical data into actionable customer profiles remains a technical challenge (Themes 1: Pre-Service Optimization and 2: Interaction Intelligence). Additionally, overly aggressive predictions or unexpected insights can be perceived as intrusive or manipulative (Themes 3 and 6), thereby undermining the trust that personalization aims to establish (Nishant, Schneckenberg, & Ravishankar, 2023).

Technology integration in service roles reshapes the autonomy, identity, and collaboration dynamics of service agents (Theme 3: service agent well-being). As tasks become increasingly augmented, service agents face shifting boundaries in decision-making and responsibility, raising concerns about professional identity and potential deskilling (Theme 6: augmentation). Studies highlight that this shift increases cognitive demands and requires service agents to continuously adapt to novel human-technology work arrangements (Gnewuch et al., 2023). Future research should investigate how training, onboarding, and interface design can facilitate adoption, how service agent system trust evolves, and how trust calibration (i.e., appropriately relying on versus questioning AI output) can be promoted. Research should focus on transparency, trust, and autonomy, while minimizing concerns about displacement and preserving the sense of purpose among service agents.

Personalized service delivery raises concerns about monitoring service agent behavior, performance, and emotional states (Themes 3: service agent well-being and Theme 4: service agent monitoring). While monitoring can benefit tailored support, it also introduces potential challenges, including increased stress and diminished autonomy. These concerns highlight the need for a comprehensive evaluation of how monitoring practices impact service agents, focusing a particular focus on human agency and transparent communication.

Building on the complexities of personalized service delivery, trust emerges as a pivotal factor shaping the relationship between service agents, technology, and organization (Theme 6: augmentation). Trust influences willingness to rely on AI systems and the perceptions of fairness, transparency, and support in hybrid work environments. Given the evolving roles and increasing human-machine interdepen-

dence, understanding how trust develops, is maintained, or erodes in personalized, technology-augmented contexts is critical. Future research should explore the mechanisms for fostering trust, considering technological design and organizational practices that align with service agents' professional values and psychological needs.

As customization deepens, it raises critical questions around data governance, transparency, and fairness (Wirtz, Kunz, Hartley, & Tarbit, 2022). The growing reliance on customer data necessitates compliance with evolving regulations, such as the GDPR and the European AI Act, particularly as personalization overlaps with high-risk categories (e.g., affective computing) (Iren, Yildirim, & Shingjergji, 2023; Kusche, 2024). This requires research on how organizations navigate regulatory landscapes and how regulations shape the design and deployment of technologies. Future work should focus on detecting and mitigating bias, as well as developing explainable AI to make processes transparent to users. These efforts are crucial for ensuring regulatory compliance, fostering user trust, and delivering equitable AI services (Nishant et al., 2023).

Finally, it is essential to underscore the distinction between technology augmentation and full automation within service environments (Theme 6: augmentation). Misunderstanding this difference can lead to theoretical blind spots, such as overlooking human expertise, or managerial missteps, including inappropriate reliance on automated systems that may erode service agent autonomy and reduce service quality. Research should examine which tasks are most appropriate for augmentation versus automation, and how these boundaries evolve as technologies and service agent expectations mature. This can help clarify how to align system capabilities with human strengths. Future research should critically examine boundaries and interactions between augmentation and automation, ensuring that technology is deployed to complement and empower human service agents rather than replace them.

2.6 Discussion

This review advances our understanding of how technology can augment rather than replace human service agents. We identify six core themes and organize them along the dimensions of technology augmentation and service customization, offering a structured lens for analyzing the evolving role of technology in service interactions. Our framework emphasizes aligning technological capabilities with human agency,

thereby ensuring the ethical, effective, and context-sensitive integration of technology. These insights lay the groundwork for future work focused on collaborative, adaptive, and human-centered service interactions.

2.6.1 Theoretical implications

This review's focus on the augmentative role of technology offers several contributions to service theory. First, it presents a human-centric frame for technology-oriented service research. Rather than conceptualizing technology as a substitute for human labor, we emphasize its role as an augmentative and co-productive force supporting service agents throughout customer interactions. This complements the dominant automation-focused narrative (e.g., M.-H. Huang and Rust (2018); Ostrom, Parasuraman, Bowen, Patrício, and Voss (2015)) and directly responds to calls for research on human-technology collaboration in service work (Larivière et al., 2017; Marinova et al., 2016). At the same time, it remains important not to overlook the potential benefits of automation. Future research should examine which tasks are most appropriate for augmentation versus automation, and how these boundaries evolve as technologies and employee expectations mature.

Second, we employ a thematic mapping approach (Clarke & Braun, 2014) structuring the literature into six core themes: pre-service optimization, interaction intelligence, service agent well-being, monitoring, emotion work, and augmentation, mapped onto the interaction process (M.-H. Huang & Rust, 2018; Zapf et al., 2003). This structure offers a theory-informed synthesis of how technologies impact not only cognitive and behavioral tasks, but also emotional and relational aspects of service work, underscoring the experiential dimensions of technological augmentation (Larivière et al., 2017).

Third, drawing on socio-technical systems theory (Trist & Bamforth, 1951) and service-dominant logic (Edvardsson et al., 2011; Vargo & Lusch, 2008), our review contributes to the understanding of how value is co-created through the dynamic interplay between humans and technology. Structuring the six themes according to the axes of customization and augmentation helps integrate the fragmented literature in this domain (Edvardsson et al., 2011; M.-H. Huang & Rust, 2018). Through the lens of socio-technical systems theory, value is an emergent outcome of tightly coupled human and technological elements. This theoretical lens emphasizes the importance of studying frontline roles not in isolation, but as embedded actors within evolving

technological systems that shape and are shaped by the behaviors of customers, service agents, and organizations (Odekerken-Schröder et al., 2022).

Fourth, by linking foundational work in computer science (LeCun et al., 2015; Russell & Norvig, 2016) with the existing service literature, we establish a cross-disciplinary foundation for understanding technological augmentation in service interactions. This conceptual synthesis offers a new lens for service researchers, focusing on the core technological capabilities that augment service agents and reshape service interactions.

2.6.2 Managerial implications

Our review highlights several practical implications for service managers. First, investing in technologies that augment, rather than replace, service agents can enhance service quality and efficiency (Theme 6: augmentation). Misunderstanding this distinction risks misguided automation strategies that undervalue human contributions. Second, managers should closely monitor service agents' well-being using real-time insights to proactively address stress and burnout, thereby reducing turnover and enhancing service agent engagement (Theme 3: service agent well-being). Third, customer interaction intelligence should extend beyond post-interaction analysis to deliver real-time insights, enabling more personalized customer interactions and better handling of challenging situations (Theme 1: pre-service intelligence; Theme 2: interaction intelligence). Fourth, performance monitoring technologies must be balanced with preserving service agent autonomy and mitigating negative psychological effects (Theme 4: service agent monitoring). Finally, transparent ethical guidelines and data management practices are crucial for fostering customer trust when deploying advanced personalization and emotional analytics (Theme 5: emotion work).

Chapter 3

Comparing neural networks for speech emotion recognition in customer service interactions

This chapter is based on an article published as Waelbers, B., Bromuri, S., & Henkel, A. P. (2022, July). Comparing neural networks for speech emotion recognition in customer service interactions. *In 2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Abstract

Automatic speech emotion recognition may assist call center service agents in deciphering and regulating customer emotions. To contribute to a successful augmentation of service agents with artificial intelligence, the main goal of this study is to identify effective machine learning approaches to classify discrete basic emotions in customer service conversations. A comparison is presented of the recognition performance of different neural network architectures on speech features extracted from service interactions in a naturalistic customer service setting. Baseline classifiers, including a zero-rule classifier, a random classifier, a frequency classifier, and non-sequential multi-class classifiers, are compared to different neural network architectures. A multi-layer perceptron, a one-dimensional convolutional neural network, and a neural machine translation model outperform the baseline classifiers, suggesting a pattern in the data relating to emotion labels. While the neural machine translation model with attention attains the highest F1-score, no significant difference in performance among the neural networks is detected. Results therefore support the use of the multi-label multi-layer perceptron as the simplest model.

3.1 Introduction

The recognition and management of emotions is one of the fundamental ingredients of the work of customer service agents in a call center service context. The constant exposure of service agents to negative customer emotions in conjunction with display rules that require the suppression of their own emotional reactions is demonstrably negatively related to performance (Goldberg & Grandey, 2007) and psychological and physical well-being (Sprigg, Stride, Wall, Holman, & Smith, 2007).

Recent work within the field of speech emotion recognition (SER) provides viable angles to alleviate service agents, both directly (e.g., providing real-time predictions of current emotional states; (Henkel, Bromuri, et al., 2020)) and indirectly (e.g., using emotions to predict perceived interpersonal stress levels; (Bromuri et al., 2021)). SER is one of the components in the domain of affective computing. Affective computing addresses the recognition and interpretation of emotions and has become a prominent field in human-computer interaction (HCI) (Picard, 1999), where the computer needs to interpret the emotion of the human to give a reply with the appropriate sentiment.

Alongside SER, emotions can be recognized by facial expressions (Bassili, 1979), body gestures (Noroozi et al., 2021), the physiological response (Shu et al., 2018), and the content of the conversation (Yoon, Byun, & Jung, 2018). Several distinct approaches to detecting emotions from speech have been suggested in prior literature. First, knowledge-based techniques are utilized frequently (Pachet & Roy, 2009). However, they require extensive knowledge of domain-specific features relevant for emotion recognition and are therefore not easily transferable across domains. To overcome this issue, a second approach relies on statistical methods, including support vector machines and Gaussian mixture models (Sailunaz, Dhaliwal, Rokne, & Alhajj, 2018). While such models demonstrate that emotions are detectable within speech, they also raise caution in their interpretation, as labels may present noise due to the arbitrariness of the task. Third, the current state-of-the-art for SER is comprised of deep neural networks, which seem to outperform standard machine learning methods (de Velasco, Justo, Antón, Carrilero, & Torres, 2018). In particular, convolutional neural network (CNN) models with different amounts of dimensions are commonly applied (Anvarjon, Mustaqeem, & Kwon, 2020; Dangol, Alsadoon, Prasad, Seher, & Alsadoon, 2020; Hajarolasvadi & Demirel, 2019; Mustaqeem & Kwon, 2021; Zhao, Mao, & Chen, 2019). However, a potential limitation of all of these approaches

is that they do not base their predictions on a sequence of emotions, but rather on single, discrete emotions without context.

Recently, the field of SER has shifted towards long short-term memory (LSTM) architectures. For instance, Wang et al. (2020) proposes a dual-sequence LSTM, where two input sequences are processed simultaneously. Another approach from Mustaqeem, Sajjad, and Kwon (2020) extracts the speech features by first adopting a key sequence segment selection and then using the resulting spectrogram of the selected sequence as input to the CNN-based feature extractor. The features are then used to train a bidirectional LSTM to learn both the temporal information and the emotional state. Xie et al. (2018) applied a bidirectional LSTM to resolve the problem of variable-length speech inputs. Finally, the contribution in Bromuri et al. (2021) also relies on a bidirectional LSTM to predict the emotion class from a sequence of speech. However, instead of predicting the emotion class of each speech snippet in the sequence, the authors solely focus on the dominant emotion per sequence.

The focus and main contribution of the underlying chapter is to evaluate categorical (rather than dimensional) models for the prediction of discrete emotions from longer sequences of spontaneous speech (cf., Ekman et al. (1987), Plutchik and Kellerman (1980)). Different models are compared to see whether additional features of the sequence or context in which the emotion appears help in recognizing and predicting the emotion label. On the one hand, the importance of the sequence of emotions is explored by using a sequence-to-sequence model. On the other hand, we compare the sequence-to-sequence model to a multi-label, multi-layer perceptron (MLP) to evaluate the presence of recurrent effects that can be exploited for the recognition of emotions, and whether the context in which the emotion appears is already sufficient to perform the prediction.

A popular approach for sequential data is neural machine translation (NMT), which is mostly deployed to translate an input language into a target language. In this chapter, an NMT architecture was used to 'translate' speech snippets into emotion labels. Since NMT relies on deep neural networks, it is faster and less memory-intensive than previous methods, such as rule-based learning (e.g., mapping every word from one language to another) and statistical machine translation (e.g., learning the probabilistic model from the data) (Castilho et al., 2017). An important application of NMT is that it is incorporated in Google Translate (Y. Wu et al., 2016). It has also been combined as an encoder and a classifier to classify emotions from textual conversations (in contrast to an encoder and decoder architecture; (Ragheb, Azé,

Bringay, & Servajean, 2019)). Bidirectional LSTM units are used in the encoder. An attention layer is added to focus on the first and last parts of the conversation of both conversation partners. Crivellari and Beinat (2020) used the NMT approach in the field of human motion trajectories. Instead of sentences, the model is fed with sequences of locations, and these motion traces were 'translated' with different trajectories.

The remainder of the chapter is structured as follows. First, Section 3.2 presents the dataset used for our experimentation, followed by a discussion of the proposed model architecture in Section 3.3. Finally, results are reported in Section 3.4, and Section 3.5 concludes with a discussion and future research directions.

3.2 Data

The dataset for this chapter comprised 363 call center service interactions among customers and service agents in a Dutch pension service context. All conversations were divided into three-second snippets, which were annotated with emotion labels. For privacy reasons, calls were annotated by the respective service agent who handled the call. Annotations were performed based on the six basic discrete emotions (i.e., anger, disgust, sadness, fear, surprise, and happiness) (Ekman et al., 1987) and a neutral class (i.e., absence of any detectable emotion, including silence). When multiple emotions were present in the same snippet, annotators were instructed to choose the most dominant emotion. The annotated emotion classes are distributed as follows: anger (2.0% of snippets), disgust (1.7% of snippets), sadness (0.5% of snippets), fear (1.2% of snippets), surprise (3.7% of snippets), happiness (4.8% of snippets), and neutral (86.0% of snippets).

Even though there is no consensus on the duration of emotions (Ekman, 1992), they frequently extend beyond several 3-second snippets (Frijda, Mesquita, Sonnemans, & van Goozen, 1991; Verduyn & Lavrijsen, 2015). We therefore hypothesized to observe a dependency between consecutive snippets and hence combined individual snippets into sequences. Here, sequences of five snippets resulted in the highest performance after a grid search with one to seven snippets in a sequence. This approach yielded 15-second snippets that were then sampled with a 25ms window and a step size of 50ms, deploying the PyAudioAnalysis (Giannakopoulos, 2015) pipeline to extract both spectral and prosodic features of the speech. The following features

were extracted: zero-crossing rate, energy, entropy/energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, mel-cepstral coefficients, chroma vector, and chroma deviation.

3.3 Methods

3.3.1 Neural machine translation

The NMT model consists of two networks: an encoder and a decoder network (Cho, van Merriënboer, Bahdanau, & Bengio, 2014). The encoder network can be any network that can handle sequential input. The decoder network predicts the translation of the input. Both are recurrent networks that can handle sequences of data by using feedback connections. Here, different architectures can be used that are suitable for detecting sequences, such as gated recurrent unit layers and LSTM layers.

The input of the decoder is the start tag, which is followed by the output of the encoder. The start tag informs the decoder that it is at the start of the sequence, after which the decoder predicts the element of the sequence that has the highest probability. When the decoder predicts the stop tag, it shows that the sequence has ended, so the decoder does not have to predict another element of the sequence. Since the goal of the decoder is to predict the next element of the sequence, the target data of the decoder is the input data of the decoder at the next time point. An attention layer has been introduced to improve the performance of the NMT (Karmakar, Teng, & Lu, 2024). This attention layer supports the decoder network in focusing on a specific part of the input (Luong, Pham, & Manning, 2015). Thus, some parts of the input have a greater impact on the output than other parts. The part of the input that is most useful for recognizing the emotion is learned by the attention model.

The underlying research examined both an NMT with and an NMT without an attention layer. The underlying architecture was an LSTM, which is a recurrent neural network that uses a memory cell in order to handle longer sequences. This memory cell contains a forget gate that decides what information is kept and what information is forgotten:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{3.1}$$

to with σ being a sigmoid activation function, W and U being the weights, b the bias,

x the input of the LSTM, and h the output of the LSTM. The encoder part of the NMT consisted of a bidirectional LSTM. In a bidirectional LSTM, the network receives the input in chronological order and in reverse order. The information from both alignments was then combined. Informed by previous and future elements of the sequence, a bidirectional LSTM can also predict the next element of the output.

The Keras library was used on top of the TensorFlow framework. The optimizer used was RMSprop, the batch size was 128, and the number of epochs for the NMT was 30 (beyond this, there is no further improvement). Since the data consisted of multiple classes, the loss function was the categorical cross-entropy, and categorical accuracy was used as a metric, which calculates the times that the prediction matches the one-hot label. Models were run with a validation split of 85% training set and 15% validation set.

Both NMT networks received five feature vectors of size 136. These feature vectors are extracted from five successive three-second audio snippets by the pyAudioAnalysis library. In line with the nature of the NMT, a $<$start$>$ element was added to the sequence. The encoder consisted of a bidirectional LSTM with 50 nodes. The nodes of both the forward part and the backward part of the LSTM were concatenated. Then, the encoder states were passed to the decoder network and used as the initial state of the LSTM. This LSTM layer contained 100 nodes and received the target sequence as the input. Subsequently, it predicted the next element of the sequence, which was a one-hot encoding of the emotion classes and the $<$end$>$ statement. The most likely emotion class was the one with the highest activation.

For the NMT model with attention (see Figure 3.1), the encoder was the same as in the no-attention model. The input of the decoder was the target sequence. The decoder contained an LSTM of 100 nodes, with the encoder states as the initial state and the target sequence as input. The output of the decoder was based on the additive attention layer that combined the encoder states and the LSTM output.

3.3.2 Baseline models and feed-forward multi-label models

To determine the performance of the NMT model, we compared it to simpler networks, including a zero-rule classifier, a random classifier, a frequency classifier, a simple multi-label CNN, and an MLP model. The zero-rule classifier always chose the largest emotion class, which was the neutral class in this case. The random classifier assigned an arbitrary emotion to every element in the sequences, where every emo-

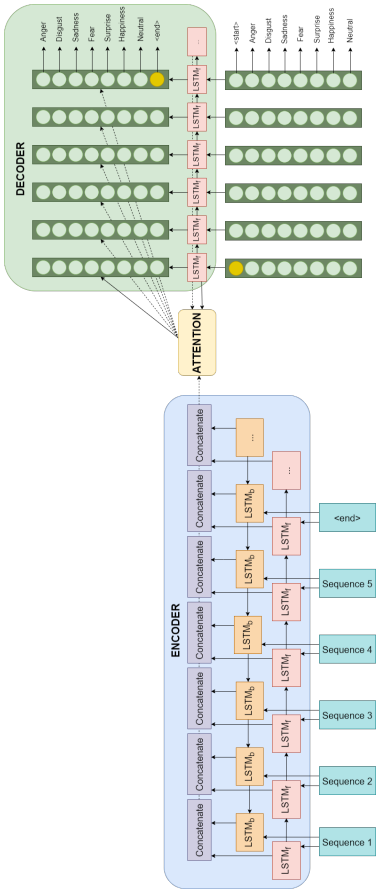

Figure 3.1: NMT architecture

tion had the same probability of getting chosen. In contrast, the frequency classifier assigned an arbitrary emotion to every element of the sequences that was selected based on a probability distribution of the training set.

CNNs consist of convolutional layers, where a kernel is slid along the input features (Goodfellow et al., 2016). In the underlying research, the CNN comprised a first one-dimensional convolutional layer with 100 nodes, a stride of two, and a ReLU activation function, followed by a maximum pooling layer. In addition, there was a second convolutional layer with 50 nodes, a stride of one, a ReLU activation function, and a maximum pooling layer. The network was flattened, and a fully-connected layer with a softmax activation function was added, resulting in the output layer with seven-node outputs for each of the five snippets in the considered sequence (7x5),

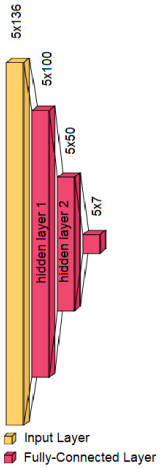

Figure 3.2: MLP architecture

representing the one-hot encoding of the emotion classes. The emotion class with the highest output value was chosen for each of the snippets.

An MLP is an artificial neural network with feed-forward layers only that are fully-connected (Wasserman & Schwartz, 1988). The MLP model consisted of an input layer, two hidden layers, and an output layer (see Figure 3.2). The input layer contained five nodes, corresponding to the number of sequences. Each input node received the 136 PyAudioAnalysis features of one item of the sequence. The first and second hidden layers had 100 and 50 nodes, respectively. Both hidden layers contained a ReLU activation function. The output layer had again a (7x5) organisation, standing for the one-hot encoding of the emotion classes and a softmax activation function. Again, the emotion class with the highest output value was chosen for each of the snippets. The rationale behind using a multi-output MLP was to evaluate whether considering the context in which an emotion occurs was sufficient to predict it (rather than also using the recurrent effects as with the NMT).

To examine the importance of the context in which each item of the sequence occurred, a non-sequential MLP and a random forest were explored. The MLP consisted of two hidden layers with 100 and 50 nodes, respectively. The output layer consisted of a seven-node output resulting in the one-hot representation of the emotion, and a softmax layer. The random forest had a maximum depth of six. The predictions of these networks were then combined into sequences of five snippets to compute the performance in a similar way as was done for the remaining models.

3.3.3 Evaluation metrics

To evaluate the models discussed above, three different performance metrics were computed for every emotion class, namely precision, recall, and the F1-score. All three scores rely on the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Precision computes the number of samples that are classified correctly in a certain class, compared to all items classified as positive for that class.

$$precision = \frac{TP}{TP + FP}$$

Recall computes the number of samples that are classified correctly in a certain class, compared to all items with the ground truth belonging to that class.

$$recall = \frac{TP}{TP + FN}$$

The F1-score computes the harmonic mean of the precision and recall scores.

$$\text{F1-score} = \frac{2 \times recall \times precision}{recall + precision}$$

These metrics are standard for machine learning frameworks. However, since our study deals with sequences, we considered the fact that multiple emotions may occur in the same sequence. When an emotion was recognized by the algorithm at any position in the sequence, it was still considered a true positive. This feature represents the assumption that the bigger context in which an emotion occurs is relevant for detecting it. For comparison, we also report the results when we instead restrict emotion recognition to the exact position of an emotion in the sequence.

3.4 Results

Different models were trained to recognize emotions from audio features, namely a zero-rule classifier, a random classifier, a frequency classifier, a multi-label CNN, a multi-label MLP, and an NMT, both with and without attention. The precision, recall, and F1-score of these models are reported in Table 3.1.

The comparisons were conducted considering two different approaches for the evaluation. In the first approach, we evaluated a classifier for detecting an emotion

in a snippet without considering where this emotion appeared in such a snippet. In the second approach, we evaluated the classifier as detecting the emotion as correct, only if it detected it in the snippet at the same position as reported in the ground truth. The reasoning behind this distinction is twofold. First, recognizing the presence of an emotion within a snippet of 15 seconds can still be useful to call center agents (Henkel, Bromuri, et al., 2020). Second, this way, we also account for situations in which an emotion is recognizable within a longer context and not only within the label of the exact three-second snippet.

Table 3.1 depicts the results of the baseline classifiers when the position of the emotion in the snippet is not considered. Table 3.2 shows the results of the baseline classifiers when the position of the emotion is relevant. The zero-rule classifier consistently labeled each segment with the most frequent emotion class (neutral), so the non-neutral classes had a precision, recall, and F1-score of zero. The random classifier and frequency classifier showed a similar performance.

Table 3.3 and Table 3.4 report the precision, recall, and F1-scores of the neural network models, with the position of the emotion not being considered and with it being considered, respectively. The more advanced models (i.e., CNN, MLP, and NMT) outperformed the zero-rule classifier, random classifier, and the frequency classifier. The NMT with attention slightly outperformed the MLP, CNN, and NMT without attention based on the F1-scores.

Table 3.5 also displays the performance of the non-sequential models applied to each three-second snippet separately, namely the non-sequential MLP and the random forest model. Comparing the sequential to the non-sequential MLP, the sequential MLP performed better (see the MLP columns in Table 3.3). The random forest model only recognized neutral and some anger, producing a similar performance as the zero-rule classifier. These results suggest that the context of the emotion is important for recognizing emotions.

Table 3.1: Baseline classifier models irrespective of position

	Zero-Rule Classifier			**Random Classifier**			**Frequency Classifier**		
	Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
Anger	0	0	0	0.05	0.59	0.09	0.04	0.10	0.05
Disgust	0	0	0	0.05	0.51	0.08	0.05	0.07	0.06
Sadness	0	0	0	0.02	0.56	0.05	0.02	0.03	0.03
Fear	0	0	0	0.03	0.51	0.06	0.03	0.07	0.05
Surprise	0	0	0	0.10	0.52	0.16	0.09	0.17	0.12
Happiness	0	0	0	0.11	0.53	0.18	0.10	0.15	0.12
Neutral	0.99	1.00	0.99	0.99	0.55	0.70	0.99	1.00	0.99

Table 3.2: Baseline classifier models with respect to position

	Zero-Rule Classifier			**Random Classifier**			**Frequency Classifier**		
	Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
Anger	0	0	0	0.02	0.16	0.03	0.01	0.02	0.02
Disgust	0	0	0	0.01	0.12	0.01	0.01	0.01	0.01
Sadness	0	0	0	0.01	0.18	0.01	0	0	0
Fear	0	0	0	0.01	0.14	0.02	0.02	0.02	0.02
Surprise	0	0	0	0.02	0.13	0.04	0.02	0.04	0.03
Happiness	0	0	0	0.03	0.13	0.05	0.04	0.04	0.04
Neutral	0.90	1.00	0.94	0.89	0.14	0.24	0.90	0.87	0.88

Table 3.3: Neural network model performance irrespective of position

	MLP			**1D CNN**			**NMT without Attention**			**NMT with Attention**		
	Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
Anger	0.12	0.17	0.14	0.03	0.50	0.07	0.18	0.10	0.12	0.21	0.09	0.13
Disgust	0.04	0.07	0.06	0.02	0.03	0.03	0.11	0.09	0.10	0.17	0.11	0.14
Sadness	0	0	0	0.02	0.91	0.04	0.09	0	0	0.03	0.01	0.01
Fear	0.09	0.10	0.09	0.03	0.88	0.06	0.07	0.05	0.06	0.08	0.08	0.08
Surprise	0.16	0.26	0.20	0.11	0.46	0.18	0.20	0.24	0.21	0.20	0.26	0.22
Happiness	0.21	0.31	0.25	0.10	0.06	0.07	0.26	0.21	0.23	0.27	0.34	0.30
Neutral	0.99	0.99	0.99	0.99	0.83	0.90	0.99	0.99	0.99	0.99	0.99	0.99

Table 3.4: Neural network model performance with respect to position

	MLP			**1D CNN**			**NMT without Attention**			**NMT with Attention**		
	Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
Anger	0.06	0.07	0.06	0.25	0.06	0.10	0.05	0.12	0.07	0.04	0.11	0.05
Disgust	0.06	0.09	0.07	0.17	0	0	0.02	0.03	0.02	0.02	0.04	0.03
Sadness	0.02	0.02	0.02	0	0	0	0	0	0	0	0	0
Fear	0	0	0	0.01	0	0	0.01	0.02	0.01	0.03	0.04	0.03
Surprise	0.04	0.05	0.04	0.09	0.04	0.08	0.06	0.06	0.06	0.08	0.06	0.07
Happiness	0.14	0.13	0.13	0.12	0.06	0.08	0.07	0.01	0.08	0.12	0.11	0.11
Neutral	0.92	0.91	0.92	0.90	0.98	0.94	0.94	0.91	0.92	0.92	0.91	0.91

Table 3.5: Model performance with no context information

	MLP			**Random Forest**		
	Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
Anger	0.05	0.16	0.08	1.00	0.01	0.02
Disgust	0.07	0.05	0.06	0	0	0
Sadness	0	0	0	0	0	0
Fear	0.03	0.05	0.04	0	0	0
Surprise	0.06	0.21	0.09	0	0	0
Happiness	0.10	0.25	0.14	0	0	0
Neutral	0.89	0.99	0.94	0.99	1.00	0.99

For all models, performance decreased when the position of the emotion was important. This suggests that the network was able to recognize an emotion in a sequence, but not at the correct time point in that sequence.

To verify that the neural networks differed from the classifiers, a 10-fold leave-one-out cross-validation was performed for the MLP, the random classifier, and the frequency classifier. The results can be found in Tables 3.4 and 3.5, showing that the MLP is significantly different from both the random and frequency classifiers, except for the neutral class. This finding may be due to the imbalance in the neutral class compared to the other emotion classes.

Leave-one-out cross-validation (100 folds) was performed for the two best-performing networks, namely the MLP and the NMT with attention. On the precision scores of these models, an independent two-sided t-test was applied to test whether the models significantly differed in their precision scores. The results of the t-test are reported in Table 3.6, Table 3.7, and Table 3.8. The p-values are larger than 0.05 for all emotions, suggesting that there is no significant difference between the MLP and the NMT with attention.

Table 3.6: Comparison MLP and random classifier

	F1-score		**Precision**	
	p-value	*t-statistic*	*p-value*	*t-statistic*
Anger	7.81 x e-14	21.40	2.22 x e-9	17.94
Disgust	4.27 x e-7	7.93	3.23 x e-8	13.76
Sadness	5.68 x e-3	3.37	4.26 x e-5	7.34
Fear	8.07 x e-10	11.84	1.03 x e-10	19.42
Surprise	1.52 x e-6	7.77	3.26 x e-10	19.39
Happiness	4.22 x e-6	7.29	1.27 x e-11	15.02
Neutral	7.68 x e-17	139.50	0.87	0.16

3.5 Discussion and conclusion

This chapter explored the sequence information in emotions during call center service interactions. MLP, NMT, and different baseline classifiers were presented to learn the emotion label corresponding to the features of a speech snippet. MLP, CNN, and NMT outperformed the baseline classifier and the non-sequential models,

Table 3.7: Comparison MLP and frequency classifier

	F1-score		**Precision**	
	p-value	*t-statistic*	*p-value*	*t-statistic*
Anger	9.20 x e-10	-17.44	2.76 x e-14	-22.64
Disgust	4.42 x e-8	-13.82	2.12 x e-7	-9.65
Sadness	4.17 x e-5	-7.29	0.02	-4.02
Fear	8.76 x e-11	-19.65	2.41 x e-10	-13.75
Surprise	3.37 x e-10	-19.43	5.97 x e-8	-9.79
Happiness	1.25 x e-11	-15.04	1.72 x e-7	-9.14
Neutral	0.99	-0.01	0.95	0.06

Table 3.8: Comparison MLP and NMT with attention

	F1-score		**Precision**	
	p-value	*t-statistic*	*p-value*	*t-statistic*
Anger	0.46	-0.74	0.92	-0.97
Disgust	0.81	-0.23	0.83	-0.22
Sadness	0.15	1.44	0.48	0.71
Fear	0.59	0.54	0.54	0.62
Surprise	0.34	0.95	0.95	-0.06
Happiness	0.27	1.12	0.95	0.06
Neutral	0.87	-0.16	0.99	0.01

indicating that a pattern between emotion class and speech features could be detected within the context in which the emotion occurred. However, this pattern did not include the recurrent information that was present in the 15s snippet of speech. Had this recurrent pattern been present, the NMT would have outperformed the MLP, since the MLP was not able to pick up recurrent information. In that case, a transformer network (J. Sun, Han, Cheng, Wu, & Wang, 2020) would have been a logical network to add to the comparison, but it was not considered here due to the lack of performance improvement by the NMT with attention. A recurrent effect extending the duration of the 15s snippet was also not captured (Verduyn & Lavrijsen, 2015). However, these longer time spans become problematic to compute for training purposes due to the large number of neurons. Additionally, these recurrent signals over a longer time span seem to be contradicted by the grid search that was performed on the number of elements in a sequence, where a sequence of five snippets performs

better than a sequence of seven snippets.

Model simplicity is crucial, especially in the case of similarly performing models. A smaller network that runs faster is preferred over a larger and slower network. The absence of statistical significance in the difference of the MLP and NMT with attention in this chapter, therefore, points to the use of the simplest model with the least amount of parameters: the MLP model. Despite simplicity being the most important criterion in previous research, future research may focus on other criteria, such as explainability. Under such circumstances, other architectures might be preferred (e.g., CNNs with a deconvolution technique (Rajwadi, Glackin, Wall, Chollet, & Cannings, 2019)).

The absence of significant differences between the MLP and NMT with attention provides a key theoretical insight into emotion recognition from speech. Complex models such as the NMT with attention and CNN are designed to exploit sequential, temporal, and context-dependent patterns, including the progression of acoustic cues over time, prosodic changes, and relationships across multiple speech snippets. In this dataset, however, these types of recurrent or temporally extended signals seem to be limited or absent, as emotional expressions are brief and relatively subtle. At the same time, the extracted speech features already capture most of the discriminative information required for classification. As a result, even simpler architectures, like the MLP, can achieve performance similar to more complex models. This suggests that in contexts with short, low-intensity emotional signals, the added capacity of complex models provides slight advantage.

In this chapter, the focus was solely on the speech features, not on the actual words spoken in the conversation. By using the linguistic information on top of the acoustic signals, both information sources can be combined to improve the emotion recognition (Yoon et al., 2018). Different types of techniques have been introduced that could serve as a promising starting point (e.g., emotional recurrent units (W. Li, Shao, Ji, & Cambria, 2022); combining reinforcement learning and domain knowledge (K. Zhang, Li, Wang, Cambria, & Li, 2022)).

The data emanated from a context with a relatively low intensity of emotional expressions. While this feature could be interpreted as a strength, it also presents a limitation: it impedes the detection of a build-up of emotions by the algorithms. Research may benefit from data in contexts that are more emotionally charged and that have both positive and negative valence. A respective two-dimensional labeling of intensity and valence may be useful due to their continuous nature. Though focus-

ing on such a binary approach may facilitate the annotation process and produce a more balanced labeling, important information on the type of discrete emotion would get lost. Therefore, a hybrid approach that combines discrete emotions and intensity may present a promising avenue for future research.

From a practice perspective, it is also not straightforward how to best evaluate model performance. While the convention is the evaluation of the F1-score as the sole performance criterion, service organizations applying the model in practice may benefit more from the model that attains the maximum precision, without considering the recall. On the level of service agents, it may be more useful to receive less frequent, yet more accurate estimates of customer emotion to assist them in navigating the emotion recognition and emotion regulation process (Henkel, Bromuri, et al., 2020). In addition, service managers may benefit more from a more reliable model as they may be more concerned with the bigger picture and the identification of extreme cases for feedback and training purposes.

Our emotion classification system detects the emotions of customers only. However, it could also be expanded to capture the emotion of service agents. Emotion recognition in conversation concentrates on the emotions expressed in an interplay between multiple speakers. Such an interplay of emotions promises to yield rich information for the understanding of emotion cycles in service interactions (Hareli & Rafaeli, 2008). Beyond its theoretical appeal, such information on the structure of the conversation could prove valuable also from a practice perspective. Information on patterns of emotion cycles could directly feed into (real-time) scripts for service agents and their further training. Eventually, systematic patterns of emotion cycles might even directly feed into the development of autonomous service bots in virtual (Ashtar, Yom-Tov, Akiva, & Rafaeli, 2021) and physical settings (for a review, see M. Blaurock, Čaić, Okan, and Henkel (2022)) with the objectives to manage role expectations and behavior in human-robot interaction in a service context (v. M. Blaurock Marah, Okan, & Henkel, 2022) and to positively shape the well-being of service consumers and agents alike (Henkel, Čaić, Blaurock, & Okan, 2020).

Chapter 4

Detecting dissatisfied customers in voice-based service interactions via multimodal AI

This chapter is based on a manuscript currently under revision (first round) at the *Journal of Interactive Marketing* in collaboration with Dr. Alexander P. Henkel and Prof. Dr. Stefano Bromuri.

Abstract

In today's customer-centric business environment, effectively identifying and addressing customer dissatisfaction is critical for sustaining long-term competitiveness. Traditional satisfaction surveys face significant limitations, including response biases and delayed feedback, which makes real-time dissatisfaction detection essential to prevent customer churn and allow immediate service recovery. This interdisciplinary research implements machine learning techniques to predict customer dissatisfaction from voice-to-voice service interactions. Textual and auditory signals from 1,144 real service interactions of a global service provider are integrated using cross-attention mechanisms to consider nuanced expressions within these interactions. Rather than simply combining multimodal data, which improves performance, this study addresses the fundamental question of how verbal and vocal components should be integrated for optimal dissatisfaction detection. Drawing on communication theory, we theorize that cross-attention operationalizes the interactivity principle according to which verbal and vocal modalities engage in interdependent, context-dependent interactions based on their proximity in place and distance. Aligning verbal tokens with temporally proximate vocal cues captures the interplay between what is said and how it is said. We show that cross-attention significantly outperforms both late-fusion and text-only baselines. These findings advance theory by linking interactivity to a concrete fusion architecture and offer a deployable method for dissatisfaction detection in real-life service settings. Furthermore, this study serves as an interdisciplinary bridge between service management and marketing on the one hand and machine learning on the other, shedding light on the diverse signal modalities and machine learning techniques essential for deciphering complex business information in customer interactions. The practical implications extend to actionable strategies for service firms, facilitating the development of effective solutions, agent training, and decision-making regarding the integration of virtual agents.

4.1 Introduction

Customer service interactions are fundamental to how businesses operate and compete in today's economy. These interactions are more than simple exchanges; they represent critical moments where customer relationships are built, maintained, or potentially damaged. Customer experience plays a central role in determining an organization's long-term success and competitive advantage (Y. Li, Xing, & Terui, 2023; Mittal & Frennea, 2010). While customer satisfaction is closely linked to retention and overall brand perception (Upamannyu & Sankpal, 2014), dissatisfaction often poses a greater risk by leading to customer churn that directly harms profitability (Zorn, Jarvis, & Bellman, 2010). Moreover, dissatisfied customers are more likely to share negative word-of-mouth (WOM), which can damage brand reputation and limit growth opportunities (Azemi, Ozuem, & Howell, 2020). The ability to detect customer dissatisfaction during real-time service interactions represents a critical opportunity for organizations to implement immediate intervention strategies. These recovery strategies can then potentially transform negative experiences into demonstrations of responsive customer care (Chang & Hung, 2018; R. Zhou et al., 2019). By addressing dissatisfaction promptly, organizations can transform negative experiences into opportunities for improvement, foster stronger customer loyalty, and improve their service processes (Aksoy, Buoye, Aksoy, Larivière, & Keiningham, 2013).

While elevating customer satisfaction and preventing dissatisfaction are key to long-term success, traditional self-reported satisfaction surveys face significant challenges that limit their effectiveness. Customer satisfaction surveys are subject to response biases like social desirability, where customers overstate satisfaction, and non-response bias, where many dissatisfied customers choose not to participate (S. Han & Anderson, 2020; K. Park, Cha, & Rhim, 2018; Stephens & Gwinner, 1998). This pattern means companies get less feedback from dissatisfied customers, even though this input is crucial for service improvements. At the same time, among customers who do leave feedback, research shows a polarization effect: responses skew toward extremes of satisfaction or dissatisfaction, while moderate opinions are relatively rare (K. Park et al., 2018; Schoenmueller, Netzer, & Stahl, 2020). In short, dissatisfaction is both underreported overall and overrepresented at the emotional extremes, leading to an incomplete and distorted picture of customer dissatisfaction. Adding to these issues, the time delay between service experience and survey response allows customers to rationalize or forget their initial dissatisfaction (S. Han

& Anderson, 2020). Moreover, dissatisfied customers often express their feelings through behaviors like negative WOM or switching providers rather than through formal feedback channels, which traditional surveys fail to capture (Stephens & Gwinner, 1998). Given these systemic limitations, direct, real-time interactions, particularly voice-to-voice communications, offer a promising alternative to detect customer dissatisfaction unobtrusively, cost-effectively, and comprehensively across all service interactions (S. Han & Anderson, 2020; Schoenmueller et al., 2020).

These voice-based interactions remain a cornerstone of customer service, with 42% of customers still preferring to speak with a human service agent (Statista Research Department, n.d.). The increasing integration of voice bots and virtual assistants further underscores the significance of voice-to-voice service interactions (Zierau et al., 2023). Yet, research on automating the detection of customer dissatisfaction by deploying artificial intelligence (AI) and machine learning (ML) techniques remains limited, particularly when considering the different communication modalities available in service interactions.

Most ML studies focus on identifying customer satisfaction from text-based information from service chats (McLean & Osei-Frimpong, 2017) or transcribed service calls (Y. Park & Gates, 2009). This focus on textual data is due to its capacity to convey detailed and comprehensive information about customer interactions. However, how we communicate also imparts meaningful information (Frick, 1985). Specific nuances may remain hidden or indiscernible in written form, and affective cues may be more perceptible in speech than in text (de Lacerda Pataca & Costa, 2023). Other aspects are entirely contingent on context and tone (e.g., sarcasm) and challenging to capture accurately in text-only models (Cheang & Pell, 2008; Y. Liu, Chi, & Sun, 2024). Despite the richness of vocal cues, previous studies have largely overlooked the potential of audio signals in voice-to-voice interactions. This leaves a vital opportunity to improve automated detection, particularly for identifying dissatisfaction, which may be more subtly conveyed through tone and prosody than through word choice alone.

The limitations of text-only approaches become particularly apparent when considering how people naturally express dissatisfaction. Critical emotional signals are often embedded in vocal characteristics like tone, pace, and prosody rather than in the specific words chosen (Banse & Scherer, 1996; Larrouy-Maestri, Poeppel, & Pell, 2024). This suggests that a more integrated approach is needed, one that can effectively combine both textual and vocal information. However, while communication

theory (such as Media Richness Theory (Daft & Lengel, 1986), Dual Coding Theory (Paivio, 1990)), and research on multimodal integration in cognitive psychology (Stein & Meredith, 1993), have extensively demonstrated that verbal and nonverbal modalities should be combined for accurate interpretation, their primary emphasis has been on establishing the benefits of such integration rather than specifying the underlying mechanisms by which this integration occurs. Thus, while these theories predict that richer channels can enhance understanding, they do not specify when or how channels should be fused for optimal impact.

Compared to Media Richness Theory and Dual Coding Theory, which predict general benefits of multiple channels, the interactivity principle predicts that the benefits of multimodality emerge when signals across channels are temporally contingent and mutually influential (Burgoon et al., 2006). It demonstrates that verbal and nonverbal modalities engage in interdependent, context-dependent interactions during human communication. Specifically, it discusses how signals interact depending on their proximity, both in place and in distance. For instance, a customer saying “that’s fine” could indicate satisfaction when spoken with a warm tone, but dissatisfaction when delivered with sarcasm or frustration, the identical words requiring entirely different interpretations based on vocal context. This interdependent, contingent interaction represents a specific mechanism for how modalities work together, rather than simply combining them. Cross-attention architectures operationalize this principle, moving beyond “more is better” assumptions to show when targeted signal integration delivers the greatest performance gains.

Building on this interactivity, our key research question becomes: how can we implement machine learning models to combine text and audio modalities to detect customer dissatisfaction effectively? According to the interactivity principle, verbal and non-verbal information can interact, compensate, and support each other, with the nature of these interactions depending on their proximity in both place and distance (Burgoon et al., 2006). This suggests that effective multimodal integration requires understanding not just what information is present in each modality, but when and where these modalities align or diverge in meaningful ways.

Current multimodal approaches in customer service analysis mostly rely on only one modality, or on basic fusion techniques that treat text and audio as independent streams (Kanchinadam, Meng, Bockhorst, Singh, & Fung, 2021; Luque, Segura, Sánchez, Umbert, & Galindo, 2017; Y. Park & Gates, 2009; Segura, Balcells, Umbert, Arias, & Luque, 2016). These methods typically concatenate features or use

simple weighting schemes, overlooking the complex, context-dependent ways vocal cues can modify or reinforce the meaning of spoken content (Luque et al., 2017). Such approaches may miss subtle but critical indicators of dissatisfaction, such as when vocal stress patterns contradict seemingly positive words, or when the timing of hesitations aligns with specific complaint-related phrases (Klasmeyer & Sendlmeier, 2000; Y. Liu et al., 2024).

Therefore, we propose using the cross-attention mechanism, which is particularly well-suited to capture these proximity-based interactions (Vaswani et al., 2017). Unlike simple fusion methods that assume modalities contribute independently, cross-attention enables models to learn dynamic, context-dependent relationships where specific textual elements attend to relevant vocal features based on their temporal and semantic proximity. This approach directly operationalizes the interactivity principle by allowing the model to identify when and where verbal and vocal cues interact, compensate for each other, or provide mutual support.

This research advances our understanding of technology-enabled customer-firm interactions in at least four ways. First, our study advances multimodal customer service analytics by examining how textual and vocal signals can be integrated to detect customer dissatisfaction in real-life service calls. Unlike prior work that largely analyzes each signal separately (Kanchinadam et al., 2021; Luque et al., 2017; Y. Park & Gates, 2009; Segura et al., 2016), we show that a cross-attention framework allows for modeling the dynamic, context-dependent interplay between what customers say and how they say it. This approach captures nuanced emotional cues that would be lost in simpler fusion methods, refining predictive accuracy while providing a deeper understanding of multimodal communication in service interactions.

Second, our work bridges communication theory and computational modeling by linking the interactivity principle with cross-attention modeling (Burgoon et al., 2006). While this principle emphasizes that verbal and nonverbal signals interact interdependently, previous studies have not linked it to specific computational mechanisms. We demonstrate how cross-attention enables models to learn when and where textual and vocal cues interact, reinforce, or compensate for each other, translating theoretical insights into a concrete algorithmic approach. This connection advances beyond theories that equate richer communication channels with better outcomes, such as Media Richness Theory (Daft & Lengel, 1986), by showing that it is the structured interactivity between verbal and vocal signals, not just their presence, that drives predictive performance.

Third, this study connects service management, marketing, communication theory, and ML by proposing a sophisticated multimodal integration technique tailored to customer service data (Lahat, Adali, & Jutten, 2015). Our findings highlight the critical role of advanced fusion models like cross-attention for capturing the complex, context-dependent dynamics of customer dissatisfaction (Y. Park & Gates, 2009; Segura et al., 2016). This contributes to expanding multimodal communication frameworks within service marketing and AI, offering actionable insights for firms aiming to leverage large-scale voice interaction data to improve dissatisfaction detection, enhance service recovery, aid organizational learning, and inform agent training programs (Libai et al., 2022; Mustak, Salminen, Plé, & Wirtz, 2021; Sieben, De Grip, Longen, & Sørensen, 2009).

Finally, our results have direct practical implications for service firms navigating evolving communication channels for customer-firm interactions. They support the adoption of voice-based customer service technologies over text-only chatbots by demonstrating the added value of vocal information for richer, more accurate customer understanding (Zierau et al., 2023). This work lays a foundation for future research on intelligent, context-aware multimodal fusion in customer service, emotion recognition, and human-computer interaction, where nuanced communication dynamics are central.

4.2 Related work

4.2.1 Customer dissatisfaction

Customer satisfaction and dissatisfaction have been important topics in service and marketing research for decades, given their central role in shaping customer loyalty, brand perception, and business performance (Dhiman & Kumar, 2022; Fraering & Minor, 2013; Terui, Hasegawa, Chun, & Ogawa, 2011). While satisfaction often leads to positive customer behaviors and long-term retention (Mittal & Frennea, 2010; Posselt & Gerstner, 2005; Ribeiro, Barbosa, Moreira, & Rodrigues, 2024; Risselada, Verhoef, & Bijmolt, 2010), dissatisfaction tends to generate stronger and more immediate reactions. Dissatisfied customers may experience intense emotions like frustration and anger (Azemi et al., 2020; Dukes & Zhu, 2019), which can lead to firm-damaging outcomes, including negative WOM (Bougie, Pieters, & Zeelenberg, 2003), brand hate (C. Zhang & Laroche, 2020), and customer churn (Ribeiro et al.,

2024).

An important insight from consumer research is that dissatisfaction is not simply the opposite of satisfaction (A. V. Lee, Moriarty, Borgstrom, & Horwitz, 2010). The two represent distinct concepts with different causes, emotional foundations, and behavioral outcomes (Bougie et al., 2003; Oliver, 1996). Some researchers even suggest that dissatisfaction is conceptually closer to the opposite of customer delight rather than satisfaction, highlighting that strategies focused solely on increasing satisfaction may not effectively prevent, detect, or address dissatisfaction (Souca, 2014). This distinction matters for both researchers and practitioners because it challenges the common practice of treating satisfaction and dissatisfaction as opposite ends of a single scale.

Recognizing this fundamental difference changes how organizations should think about dissatisfaction, from simply the absence of positive experiences to a critical business priority that requires dedicated attention. Effective dissatisfaction management not only helps avoid high-impact risks but can also create positive outcomes. Research shows that organizations can turn dissatisfied customers into satisfied ones through effective service recovery (Swanson & Kelley, 2001), and customer feedback from negative experiences can drive meaningful improvements in service delivery (Andreassen, 1999; Lapré, 2011). Given the significant consequences of undetected dissatisfaction and the limitations of satisfaction-focused approaches in identifying it, developing methods specifically designed to detect and address dissatisfaction becomes essential.

4.2.2 Customer (dis)satisfaction detection

Given the significant impact of customer dissatisfaction, the ability to identify it accurately during service interactions represents a critical capability for both researchers and practitioners. Contact centers provide particularly valuable settings for this type of detection because they handle large volumes of customer interactions daily, creating rich data streams that can reveal customer emotional states. These data sources include conversation topics (Papadia et al., 2022), customer intent (Zhong & Li, 2019), and call summaries (Tamura, Ishikawa, Saikou, & Tsuchida, 2011), along with vocal indicators such as emotional tone (Ashtar et al., 2023), conversational fluency (Fernández-Sabiote & López-López, 2020), emotional states (Waelbers, Bromuri, & Henkel, 2022), and overall sentiment (Grljević & Bošnjak, 2018). Beyond understand-

ing current interactions, these signals can also help predict future customer behaviors like churn (Zhong & Li, 2019) or responsiveness to sales opportunities (Bailey & Clark, 2007).

The application of ML technology is rapidly gaining momentum in the field of customer service. Here, many studies focus on predicting customer satisfaction from speech-to-text transcripts, drawing on emotional and semantic cues found in customer-agent conversations (Ashtar et al., 2023). Researchers have applied various ML models to this challenge, including support vector machines (Y. Park & Gates, 2009), ranking models (Bockhorst, Yu, Polania, & Fung, 2017), joint learning frameworks (Ando, Masumura, Kamiyama, Kobashikawa, & Aono, 2017; Ando et al., 2020), graph neural networks (Kanchinadam et al., 2021), and approaches that use soft labels (Manderscheid & Lee, 2023). Table 4.1 provides a comprehensive overview of studies predicting customer satisfaction in service interactions. This diverse research demonstrates significant progress in detecting satisfaction levels across different industries and data types.

However, this emphasis on text analysis overlooks a fundamental reality of customer service: voice communication remains an essential channel for customer service, with 64% of customers considering it the most effective one (Statista Research Department, n.d.). As automation expands and voice assistants like Siri and Alexa become more common (Hoy, 2018), voice-based customer service interactions are becoming increasingly important (Forbes & Ravinutala, 2023; OpenAI, 2023). Voice signals contain rich emotional information beyond word content, allowing detection of internal states through vocal characteristics like tone and inflection (Bromuri et al., 2021; Frick, 1985). Recent research has begun exploring these voice-based cues to improve satisfaction detection models (Zierau et al., 2023). A customer might express satisfaction with their words while their vocal tone reveals underlying frustration, or conversely, their tone might indicate acceptance despite negative semantic content. Despite the availability of rich information in call center conversations, most satisfaction detection research continues to rely solely on textual features, potentially missing crucial emotional signals embedded in vocal characteristics (Zierau et al., 2023).

Table 4.1: Literature overview

Author (Year)	Data Type	Text	Audio	Method	Evaluation Metric	Key Findings
Ando et al. (2017)	Transcripts of acted call center conversations	X		Joint Modeling Long Short-Term Memory Recurrent Neural Network	Macro F1-score	Joint modeling shows better predictive accuracy than support vector machine.
Ando et al. (2017)	See Ando et al. (2017)	X		See Ando et al. (2017)	Accuracy, Macro F1-score	Joint modeling outperforms conventional models in both turn- and call-level estimations.
Baier, Kühl, Schüritz, and Satzger (2021)	IT incidents encounters	X		Logistic Regression; Gaussian Naïve Bayes; Support Vector Machine; Random Forest; Multilayer Perceptron	F2-score	Logistic regression shows F2-score of .379, outperforming traditional regression and classification methods.

Author (Year)	Data Type	Text	Audio	Method	Evaluation Metric	Key Findings
Bockhorst et al. (2017)	IT incidents encounters	X		Ranking Model; Convolutional Fitting Function	Pearson correlation, Spearman correlation, mean absolute error	The ranking model with convolutional fitting outperforms conventional approaches.
(Bromuri et al., 2021)	Call center recordings		X	Long Short-Term Memory Recurrent Neural Network	Accuracy	The classifier shows a balanced accuracy of 68%.
Kanchinadam et al. (2021)	Transcripts of incoming call center calls	X		Graph Neural Networks	Spearman correlation, precision@k	GNNs outperform conventional models.
Ko, Hsu, Liu, and Yang (2022)	Audio clips in which participants indicate their satisfaction levels in Mandarin		X	Support Vector Machine; Autoencoder and Long Short-Term Memory-Recurrent Neural Network	Accuracy	Prosodic and MFCC features are combined and fed into an SVM and an LSTM-RNN. The SVM with autoencoder outperforms the LSTM-RNN with autoencoder (79.31% vs. 75.86% accuracy).

Author (Year)	Data Type	Text	Audio	Method	Evaluation Metric	Key Findings
Manderscheid and Lee (2023)	Transcripts of incoming call center calls	X		Transformer-based Big Bird Model	Precision, Recall	Classification was first binary, then restored to five output classes using probability thresholds. Soft labels outperform hard labels.
Y. Park and Gates (2009)	Customer satisfaction survey notes and call transcripts	X		Decision Tree; Naïve Bayes; Logistic Regression; Support Vector Machine	Accuracy, Precision, Recall, F1-score	In 5-scale satisfaction, SVM shows best performance (66% accuracy), and decision tree for 2-scale satisfaction (89% accuracy).
Parra-Gallego and Orozco-Arroyave (2022)	IEMOCAP, RAVDESS, KONECTADB corpus (voicemails with spontaneous agent evaluation)		X	x-vectors and i-vectors; Disvoice Framework	Specificity, Recall, ROC	I2010PC feature set works best for acted datasets, while articulation features work best in real-world call center data.

Author (Year)	Data Type	Text	Audio	Method	Evaluation Metric	Key Findings
(Segura et al., 2016)	French political debates; Spanish phone call conversations		X	Convolutional Neural Networks	Recall, Correlation coefficient	Using raw audio as input for a CNN performs similarly to traditional methods such as Mel filter banks, without pre-processing.
The underlying study	Call center recordings plus corresponding transcripts	X	X	Long Short-Term Memory-Recurrent Neural Networks; Cross-Attention	ROC-AUC	Combining auditory and textual features using cross-attention outperforms separate modalities and models without cross-attention.

This limitation becomes even more problematic when considering that while customer satisfaction studies typically include dissatisfaction as a category, dissatisfaction detection as a primary research objective remains underexplored. While satisfied customers may continue customer-firm relationships passively, dissatisfied customers actively drive churn, negative WOM, and revenue loss, making their early identification paramount for business sustainability. The limited research specifically targeting dissatisfaction detection has focused on narrow contexts: text mining of hotel reviews (X. Xu & Li, 2016), automotive industry predictions using traditional ML (Meinzer, Jensen, Thamm, Hornegger, & Eskofier, 2016), telecommunications complaint classification (Lukitasari & Hidayat, 2020), and internet application performance issues (Joumblatt, Chandrashekar, Kveton, Taft, & Teixeira, 2013). Only one study implemented dissatisfaction detection based on real-life call center conversations (Cong et al., 2016). Consequently, we lack robust, generalizable methods for detecting customer dissatisfaction from real-life interactions, particularly those that leverage multimodal data combining both text and audio.

Current research methodologies further limit practical applicability, with studies predominantly using either artificial, acted dialogues (Parra-Gallego & Orozco-Arroyave, 2022) or genuine customer exchanges in controlled settings (Ko et al., 2022; Parra-Gallego & Orozco-Arroyave, 2022). At the same time, few examine authentic call center interactions that capture the spontaneous complexity of real-life customer expressions (Segura et al., 2016). Moreover, most studies focus solely on accuracy or F1-scores, overlooking the inherent class imbalance in customer satisfaction data, where most interactions indicate satisfaction or neutrality. This methodological oversight severely limits practical utility for service businesses that specifically need to identify the minority class, dissatisfied customers, who drive disproportionate business impact.

To address these critical gaps, this chapter introduces the cross-attention mechanism from computer science to effectively combine textual and vocal features for dissatisfaction detection in real contact center interactions. In this research, we focus on how to combine text and audio signals. While previous approaches have treated text and audio as separate information streams or used simple concatenation methods, our cross-attention architecture enables dynamic interaction between modalities, allowing the model to identify when vocal emotional indicators enhance, contradict, or provide additional context to textual content. Additionally, we evaluate model performance using ROC-AUC (Naidu, Zuva, & Sibanda, 2023), which repre-

sents the model's capability to distinguish between dissatisfied customers and satisfied customers across all possible classification thresholds, providing comprehensive insight into the decision-making trade-offs faced by service firms and enabling practitioners to balance competing objectives of identifying dissatisfied customers while minimizing false positives.

4.3 Methodology

4.3.1 Data

A leading international logistics service provider compiled a dataset of 1,240 voice-based, human-to-human customer service calls. These calls were randomly sampled from the company's worldwide English-language interactions between consumers and service agents in the first quarter of 2023. The data were collected exclusively from inbound, locally handled calls (i.e., not routed to offshore service centers). The callers were customers seeking assistance with questions, placing orders, or resolving problems related to logistics services.

After a thorough pre-screening to ensure language consistency, 96 calls were excluded as they (partly) contained conversations in languages other than English, resulting in a final dataset of 1,144 English-speaking interactions. These calls provided the basis for statistical analysis and audio feature extraction. The company supplied anonymized and diarized transcriptions of these calls, with diarization referring to the process of attributing each segment of dialogue to either a service agent or a customer.

To predict customer dissatisfaction, all interactions were annotated by one of two expert coders with extensive expertise in customer interaction analysis. This approach was chosen because of the absence of systematic customer dissatisfaction scores in the company's database. At the same time, this approach avoids well-documented limitations of customer satisfaction surveys, such as self-selection bias (Giovanna & Luciana, 2011), response bias (S. Han & Anderson, 2020), and low response rates that result in smaller, potentially unrepresentative samples (S. Yang & Kruschke, 2024). By relying on expert coders, we addressed these challenges and provided a direct and systematic assessment of customer dissatisfaction. To ensure reliability and validity, coders participated in regular calibration sessions to align their interpretations, and all ambiguous cases were reviewed collaboratively to reach con-

sensus. This process not only enhanced consistency but also ensured a robust and replicable coding procedure.

Conversations were classified into two different categories: DSAT and SAT. Conversations were labeled dissatisfied (DSAT) when the customer expressed frustration, disagreement, dissatisfaction with the solution, or when the issue remained unresolved. All other interactions, where customers are neutral or express satisfaction, are grouped as SAT and will be referred to as satisfied for the remainder of the chapter. However, this category includes both satisfied and neutral conversations.

This binary classification is beneficial in the logistics context, where customers rarely express high levels of satisfaction or happiness. Because very satisfied or highly positive interactions are infrequent, applying a more granular scale may lead to inconsistent labeling at the higher end. The 2-point scale reduces ambiguity, improving annotation consistency by focusing on whether dissatisfaction is present or not.

This annotation process resulted in the identification of 943 satisfied and 201 dissatisfied customer interactions. Example sentences for both categories can be found in Table 4.2. The dataset was subsequently divided into training, validation, and testing sets as detailed in Appendix A.

Table 4.2: Sample illustrations of service interactions

Satisfied Service Interactions:
"Nothing else, thank you so much for your help." "It is already with the driver? That is great!" "Perfect. Have a great evening."
Dissatisfied Service Interactions:
"No, I just told you I was at home! You're not listening!" "The driver was very rude to me!" "Because there is no point here! You are saying the same thing every day, right?"

4.3.2 Exploratory data analysis

Audio

The dataset comprises 1,144 calls, with call durations averaging 231 seconds (SD = 144 seconds). The data ranges from 24 seconds to 32 minutes (we recall 1,892 seconds from Appendix A). Calls resulting in satisfied customers had an average

duration of 225 seconds (SD = 138 seconds), ranging from 24 to 1,892 seconds. In contrast, calls from dissatisfied customers had an average duration of 261 seconds (SD = 169 seconds), with these calls ranging from 55 seconds to 1,512 seconds in length. Calls resulting in satisfied customers had a significantly lower duration than calls resulting in dissatisfied customers (Mann-Whitney U test U(df) = 109,460.5 (200), $p < 0.001$).

4.3.3 Text

Within the 1,144 calls, an average of 93 sentences was spoken (Appendix A). The diarization process during call transcription distinguishes each sentence as either spoken by the caller or the service agent. Thus, we can calculate the respective sentence counts for each conversation partner (Appendix B). The total number of sentences spoken in a conversation did not differ significantly (U(df) = 94,623.5 (200), $p > 0.05$). However, the number of words spoken in conversations labeled as satisfied was lower than in the dissatisfied ones (U(df) = 114,541.0 (200), $p < 0.001$), explaining the difference in duration of the conversation.

4.3.4 Conversational features

General conversational features are presented in Table 4.3. A logistic regression analysis was performed, revealing that none of these features were significant predictors of satisfaction, and therefore, they will not be considered for further analysis.

4.3.5 Models

Our study employs various methodologies to detect customer dissatisfaction. Initially, we explore the efficacy of a custom neural network model, trained on a designated subset of data known as the training set. The model's robustness is then assessed on discrete and non-overlapping data subsets, commonly referred to as the validation set and test set. As an alternative, we also consider the application of an established pretrained model, which has been trained on a similar dataset and is subsequently adapted to our unique data corpus. In our research, we focus on developing distinct models for audio, text, and their combination. Additionally, we integrate the TweetEval pretrained model, which has been trained on tweet sentiment analysis, into our study (Barbieri, Camacho-Collados, Neves, & Espinosa-Anke, 2020).

Table 4.3: Conversational features

Conversational Feature	**Description**	**Measurement Unit**	**Mean (Std.)**
Duration	Length of the conversation	Seconds	231.3 (144.9)
Agent percentage	Percentage of sentences spoken by the agent	Ratio	61.2 (14.3)
Caller percentage	Percentage of sentences spoken by the caller	Ratio	38.8 (14.3)
Agent number of sentences	The number of sentences spoken by the agent	Sentences	50.9 (22.8)
Caller number of sentences	The number of sentences spoken by the caller	Sentences	42.4 (24.5)
Total number of sentences	The number of sentences spoken in total	Sentences	93.2 (42.5)
Agent number of words	The number of words spoken by the agent	Words	351.4 (156)
Caller number of words	The number of words spoken by the caller	Words	246.1 (150.5)
Total number of words	The number of words spoken in total	Words	597.5 (265.3)

Custom models

Neural networks, a subset of ML, utilize artificial neuron interconnections to recognize patterns in data. Recurrent neural networks (RNNs) are specific neural networks that leverage the temporal sequence of data to focus on the most essential parts of a signal. The latter are particularly suited for identifying customer dissatisfaction as they can implicitly learn the parts of the conversation that are most indicative of customer dissatisfaction, such as the peak and end affective displays (Ashtar et al., 2023). Given the sequential nature of service interactions and conversation durations of up to 31 minutes in our dataset, Long Short-Term Memory (LSTM) networks are implemented to preserve information over extended periods. LSTM helps remember valuable information of the signal by storing and accessing information over longer time spans. This chapter involves training several LSTM-RNN models, tailored for audio, text, their combination, and the integration of audio with the pretrained model (See Figure 4.1). The implementation specifics are detailed in Appendix C and Appendix D.

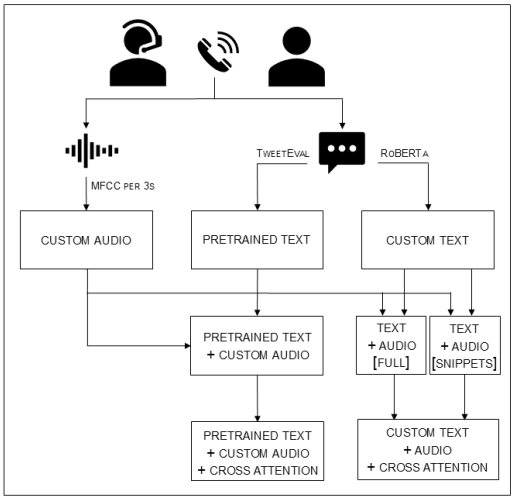

Figure 4.1: Model overview

Audio

Mel Frequency Cepstral Coefficients (MFCC), which capture the short-term power spectrum of sound, are used for audio feature extraction (Abdul & Al-Talabani, 2022). Audio content is split into three-second segments, from which MFCC are derived using the Python *librosa* package. These coefficients are then fed into an LSTM model. Additionally, Whisper embeddings were extracted to predict customer dissatisfaction. We incorporated the Whisper base version, a generative speech-to-text model trained on a vast corpus of 680,000 hours of labeled audio data and capable of generating transcripts in around 100 languages (Radford et al., 2023). Given that the Whisper embeddings did not outperform the simpler MFCC embeddings in our evaluation, and considering the principle of favoring simplicity, we have decided not to proceed with the Whisper model in subsequent steps.

Text

For text analysis, we exclusively consider customers' sentences, as our objective is to discern dissatisfaction from the customer's perspective. Observations indicate that the service agent's sentiment is typically neutral or positive, likely a result of emotion display rules at work and professional training, reducing the model's per-

formance[1] (L. Lee & Madera, 2019). To extract textual encodings, we employ an advanced BERT model, RoBERTa, which demonstrates superior performance over other BERT variants (Cortiz, 2021; Devlin, Chang, Lee, & Toutanova, 2019). These encodings provide the input for the LSTM-RNN model.

Audio and text

The combination of text encodings with audio MFCC forms the input for the LSTM model. We explore two methodologies: one combining the complete text transcript encodings with audio features, and the other combining text snippet encodings with audio snippet features.

Pretrained model

Pretrained models are deep learning models trained on large datasets. They can recognize general patterns and features that can then be transferred to similar data (S. Han & Anderson, 2020). These models offer the benefits of time and resource efficiency, as they can be directly applied or minimally fine-tuned to a new dataset. However, finding a model that generalizes effectively across different types of datasets is challenging. Our study leverages the TweetEval framework, where the RoBERTa sentiment model has been fine-tuned on a corpus of approximately 58 million Twitter messages from the TweetEval benchmark to classify sentiments as positive, negative, or neutral (Barbieri et al., 2020). RoBERTa, an optimized iteration of Google's BERT model, differs from BERT in its enhanced training data and computational power, an absence of next sentence prediction, and a modified masking strategy (for more differences, see Barbieri et al. (2020)).

Given its training on short Twitter posts, we split call transcripts into smaller segments for sentiment labeling by RoBERTa as positive, neutral, or negative, leading to an aggregate conversation score. Since our primary focus lies in distinguishing dissatisfied customers, we merge the scores for positive and neutral sentiments for our classification purposes in line with our data annotation procedure. Note that the pretrained model's training was not on customer dissatisfaction per se but on sentiment. Yet, these constructs are closely related, making the model well-suited for our purposes (Y. Kim, Levy, & Liu, 2020). The advantage of a pretrained model lies in

[1] Models incorporating only customer text signals were found to outperform those using combined customer and agent text signals

its foundational training on expansive datasets, making extensive model training and tuning by using limited specialized data obsolete.

Pretrained models and audio

We examine the efficacy of combining pretrained model outputs with audio features to ascertain whether this fusion enhances the detection of customer dissatisfaction.

Cross-attention

Beyond the custom-trained models, we implement a cross-attention mechanism. Attention is an algorithmic focus akin to a cognitive spotlight that automatically focuses on the important parts of a data sequence (Vaswani et al., 2017). When merging two signals, such as audio and text in our case, we apply cross-attention to blend these signals most optimally by identifying and emphasizing their most essential connections (Gheini, Ren, & May, 2021). This blending is related to the interactivity principle discussed earlier, where signals interact depending on their proximity in space and time. In essence, this technique assigns greater weight to specific text segments based on corresponding audio cues (for further technical elaboration, please refer to Appendix E). We train two models with a cross-attention layer, each sharing a similar architecture to our custom models: the first combines RoBERTa text embeddings with audio MFCC, and the second merges TweetEval pretrained model embeddings with audio MFCC.

Evaluation metrics

The threshold for classifying conversations as dissatisfied can be adjusted to align with company objectives, influencing the distribution of prediction classifications and the resulting confusion matrix metrics (see Table 4.4). Although an ideal threshold would perfectly categorize all conversations, such an outcome is improbable, requiring a strategic choice: Companies aiming to identify dissatisfied customers might opt for a lower threshold, accepting a higher rate of false positives. This ensures that a large share of dissatisfied customers is correctly identified, offering service recovery opportunities. However, at the same time, the model will flag more false positives, that is, customers who are labeled dissatisfied even though they are not.

This approach may be costly for service firms in terms of financial and temporal resources. Conversely, a higher threshold would prioritize the identification of satisfied customers at the expense of missing dissatisfied ones, which could have significant repercussions if service recovery opportunities are overlooked. This delicate balance between misclassifying satisfied customers as dissatisfied and vice versa is critical. Considering the dataset's imbalanced nature (i.e., a 17% dissatisfaction rate), we require a performance metric that provides information about both satisfied and dissatisfied classes. We therefore adopt the Receiver Operating Characteristics (ROC) Area Under the Curve (AUC) (Bradley, 1997). This metric, illustrating the trade-off between the False Positive Rate (FPR) and True Positive Rate (TPR, also referred to as recall or sensitivity), provides a nuanced understanding of a model's discriminatory power across various thresholds between 0 and 1. TPR and FPR are computed by using the values of the confusion matrix (see Table 4.4), with Equation 4.1 and Equation 4.2, respectively. Model assessments and comparisons are based on distinct TPR-FPR pairs and their overall AUC performance.

$$\text{TPR} = \frac{TP}{TP + FN} \tag{4.1}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{4.2}$$

Table 4.4: Confusion matrix

	Predicted SAT	**Predicted DSAT**
Actual SAT	TN	FP
Actual DSAT	FN	TP

Note: The columns indicate the predictions from the model. The rows indicate the labels provided by the company. SAT stands for satisfied, DSAT for dissatisfied. TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative.

4.3.6 Audio

The audio-based models employing MFCC achieved an AUC of 0.55, as depicted in the light grey line with circles in Figure 4.2. Given that an AUC of 0.50 equates to chance-level performance, the audio model demonstrates minimal predictive benefit beyond random classification. This suggests an equal likelihood of correctly identify-

ing a dissatisfied customer and incorrectly labeling a satisfied customer as dissatisfied.

4.3.7 Text

Analysis of the text model[2], which was trained on RoBERTa-generated embeddings, revealed an AUC of 0.75 (the light grey line with squares in Figure 4.2). This outcome indicates a substantial capability of the model to differentiate between satisfied and dissatisfied customer interactions.

4.3.8 Audio and text

In assessing the combined audio-text models, the full transcript encoding approach yielded an AUC of 0.58 (the light grey solid line in Figure 4.2). In contrast, the sequential snippet-based encoding model attained an AUC of 0.69 (the light grey dotted line in Figure 4.2).

4.3.9 Pretrained model

The pretrained TweetEval model, visualized in the dark grey solid line in Figure 4.2, along with its integration with MFCC audio features, represented in the dark grey dotted line, both achieved an AUC of 0.78 on the test set. Each of these three configurations exhibited robust performance in the detection of customer dissatisfaction in absolute terms. The pretrained models demonstrated nuanced differences in performance across various thresholds, with the integration of audio features appearing to enhance the identification of dissatisfaction at higher TPR yet introducing reduced performance at lower TPR rates.

4.3.10 Cross-attention

The cross-attention model merging MFCC audio features with RoBERTa text embeddings achieved an AUC of 0.83, illustrated in the black solid line in Figure 4.2. The combination of pretrained TweetEval text embeddings with MFCC audio features and cross-attention resulted in an AUC of 0.84, denoted in the black dotted line in Figure

[2]Models incorporating only customer text signals were found to outperform those using combined customer and agent text signals.

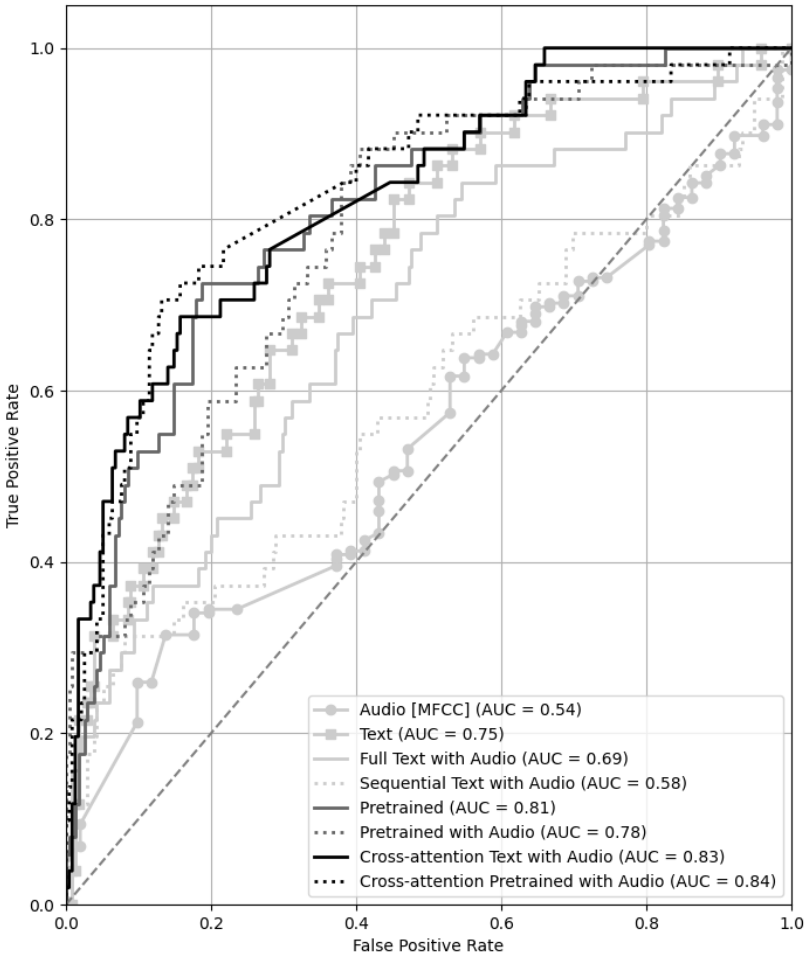

Figure 4.2: Model overview

4.2. As depicted in Figure 4.2, the latter model exhibited a significant improvement over the former within an FPR range of 0.2 to 0.65. In contrast, the former was superior at the extremities of the FPR spectrum.

4.3.11 Model comparison

A comparative analysis between every model pair was conducted using 5-fold cross-validation to validate their distinctiveness. Specifically, it was assessed whether the models significantly differ in their training set predictions. The results can be found in Table 4.5. Only two model pairs had a non-significant result, namely the audio model (the light grey line with circles Figure 4.2) was not significantly different ($p > 0.05$) from the pretrained model combined with audio features (the dark grey dotted line in Figure 4.2) and the RoBERTa text model (the light grey line with squares in Figure 4.2) was not significantly different from the RoBERTa features combined with the audio features (the light grey solid line in Figure 4.2). All other models were significantly different from each other ($p < 0.05$).

The findings highlight clear patterns in how different data types and modeling approaches affect performance. Models using only audio data perform at a level close to random guessing, showing that audio features alone contribute little predictive value. When combining audio and text in a conventional model, the results improve to a reasonable level. Importantly, using full-text encodings outperforms encoded text snippets, indicating that comprehensive text representations provide more robust inputs for prediction tasks. Among text-only models, pretrained models consistently outperform custom models, underscoring the effectiveness of leveraging large-scale pretraining for textual features. Adding audio data to pretrained text models without further integration does not enhance performance, suggesting that naïve combination strategies are insufficient to exploit complementary information in audio signals. However, incorporating audio with text in both pretrained and custom models using cross-attention mechanisms leads to significant performance gains in predicting customer dissatisfaction. These findings emphasize that the integration of multimodal signals requires an appropriate methodological approach to realize their synergistic potential fully. The superiority of the cross-attention model over simpler multimodal combinations aligns with previous work about the interactivity principle (Burgoon et al., 2006), affirming that effective signal fusion is essential for improving predictive performance.

Table 4.5: Comparative analysis of model pairs: 5-fold cross-validation testing for significant differences in training set predictions

	Audio	**Text**	**Full Text with Audio**	**Seq. Text with Audio**	**Pre-trained with Audio**	**Pre-trained**	**Cross Att. RoBERTa with Audio**	**Cross Att. Tweet-Eval with Audio**
Audio	x	< 0.001	< 0.001	< 0.001	> 0.05	< 0.001	< 0.001	< 0.001
Text	< 0.001	x	> 0.05	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Full Text with Audio	< 0.001	> 0.05	x	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Seq. Text with Audio	< 0.001	< 0.001	< 0.001	x	< 0.001	< 0.001	< 0.001	< 0.001
Pre-trained with Audio	> 0.05	< 0.001	< 0.001	< 0.001	x	< 0.001	< 0.001	< 0.001
Pre-trained	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	x	< 0.05	< 0.001
Cross Att. RoBERTa with Audio	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.05	x	< 0.001
Cross Att. Tweet-Eval with Audio	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	x

4.4 Discussion

This chapter represents a contribution to the integration of ML techniques within the domain of detecting customer dissatisfaction in voice-to-voice service interactions. Our investigation into the predictive power of verbal and vocal cues for customer dissatisfaction has added evidence to the complexity and dynamism inherent in customer interactions, particularly in identifying the indicators that traditional surveys often miss (Ashtar et al., 2023). In particular, we examine how verbal and vocal signals can be integrated through cross-attention to allow the model to dynamically weigh and align relevant information from each modality, providing a principled approach to multimodal fusion rather than treating each input independently.

The comparative analysis of neural network models has revealed that while the audio model alone shows limited predictive capability for dissatisfaction detection, its integration with text data through non-sequential modeling significantly enhances performance. The non-sequential model, which considers the conversation in its entirety rather than in segmented parts, aligns with the observed variability of customer dissatisfaction throughout a service interaction. This approach is particularly valuable given that dissatisfaction may manifest subtly through vocal characteristics like tone, pace, and prosody rather than through explicit verbal complaints (Banse & Scherer, 1996; Larrouy-Maestri et al., 2024). The implementation of cross-attention mechanisms has proven to be especially effective, indicating that the complex interplay of verbal and vocal signals can be better understood through such integrative approaches (Vaswani et al., 2017).

Our findings provide empirical support for the interactivity principle as outlined by Burgoon et al. (2006), reinforcing the idea that verbal and non-verbal cues do not function in isolation but rather interact to enhance the overall communicative effect. Specifically, the superior performance of models utilizing cross-attention between audio and text features, as compared to models without cross-attention, highlights the critical role of interactivity in predicting customer dissatisfaction. This suggests that the temporal and contextual alignment of verbal and non-verbal signals, proximity in both place and distance, can significantly enhance the interpretative power of these features for dissatisfaction detection, even in an ML context.

Moreover, our study extends on how to combine multimodal signals by demonstrating that not only does the integration of these signals improve predictive outcomes, but the method of integration is equally crucial (Burgoon et al., 2006). The

cross-attention mechanism facilitates a more nuanced interaction between modalities, allowing for a richer, more contextually relevant synthesis of information that can capture the subtle emotional signals often missed by text-only approaches. This finding is particularly significant for service organizations seeking to implement real-time intervention strategies, as it enables the detection of dissatisfaction before it escalates to formal complaints or customer churn (Borah, Prakhya, & Sharma, 2020; Risselada et al., 2010). Our results suggest that future research and applications in customer service should prioritize not just the inclusion of diverse signal types but also their intelligent combination based on their temporal and spatial alignment to capture the complex dynamics of human communication fully.

4.4.1 Theoretical implications

The current research delves into the complexity of human communication, particularly emphasizing the non-verbal, vocal elements in voice-to-voice service interactions that are critical for detecting customer dissatisfaction (Segura et al., 2016). Our findings directly address the limitations of traditional satisfaction surveys by demonstrating how vocal cues can reveal dissatisfaction that customers might not explicitly express or that may be missed due to response biases (S. Han & Anderson, 2020; K. Park et al., 2018). This chapter reaffirms that within customer service, vocal nuances, particularly those that escape transcription, such as tone, pitch, and intonation, contain crucial information beyond the textual characteristics. Furthermore, the chapter expands the boundaries of communication theory into the digital realm, making the different layers of communication detectable with the help of ML, and suggesting its relevance also for the management of service interactions.

The theoretical implications of our research suggest a shift in the understanding of customer dissatisfaction in voice-based service interactions. The study's integration of audio and text through cross-attention mechanisms provides a framework that aligns with the multifaceted nature of human interactions and thereby challenges predominantly unimodal, text-centric models of customer satisfaction detection (Y. Park & Gates, 2009). Our approach operationalizes the interactivity principle by demonstrating how cross-attention mechanisms can capture the context-dependent relationships between verbal and vocal cues that are essential for accurate dissatisfaction detection (Burgoon et al., 2006). Voice carries not only the semantic content of words but also information on paralinguistic cues, such as tone, pitch, and in-

tonation, which play a crucial role in conveying emotions and attitudes, which are encapsulated in the MFCC features (Larrouy-Maestri et al., 2024; Yildirim & Iren, 2023). This is particularly relevant for dissatisfaction detection, as customers may use neutral or positive words while expressing frustration through vocal tone, creating the type of interdependent, contingent interactions that require more complex multimodal analysis. Employing ML to decode the complex interplay of linguistic and acoustic features, our research contributes to the understanding of voice-based service interactions. The multimodal approach provides empirical evidence supporting the theoretical claim that auditory cues play a critical role in human communication, particularly in conveying affective states and attitudes that are central in determining customer dissatisfaction (Grewal, Herhausen, Ludwig, & Ordenes, 2022).

Resulting implications also extend to the future application of ML for customer service, supporting the consideration of the auditory component of human interaction. With an increasing prevalence of AI-driven service agents in the form of bots, the automated assessment of customer dissatisfaction becomes crucial (Lu et al., 2020). Here, voice bots may be preferred over chatbots, as the auditory signal can help companies in detecting customer dissatisfaction. Further, the effectiveness of voice bots will increasingly depend on their ability to process and respond to the subtleties of human emotion conveyed through voice, thus reinforcing the need for sophisticated multimodal AI models (Grewal, Herhausen, et al., 2022).

Finally, our research connects service management, communication research, and ML, emphasizing the importance of integrating diverse signal modalities for comprehensive business insights in a technical way supported by communication theory (Y. Park & Gates, 2009; Segura et al., 2016; Verma, Agrawal, Patel, & Patel, 2016). By bridging the interactivity principle with computational methods, we demonstrate how theoretical insights about multimodal communication can be practically implemented for real-time dissatisfaction detection. The use of ML techniques, such as BERT and cross-attention models, to analyze verbal and vocal cues from customer interactions advances our understanding of customer dissatisfaction. It shows how complex, emotion-laden communication can be quantified and interpreted, contributing to the intersection of computer science and service management (Ameen et al., 2011; Mustak et al., 2021). These insights underscore the potential for actionable strategies that service firms can employ to enhance both effectiveness and efficiency in assessing and managing customer dissatisfaction, enriching the interplay between human-centric service and technology-driven solutions, enabling proactive service

recovery and intervention strategies (Lervik Olsen, Witell, & Gustafsson, 2014).

In a world that is increasingly reliant on voice-to-voice service interactions, whether between humans or between humans and voice bots, the theoretical implications of our research are positioned to influence a host of applications, from customer service protocols to the design of empathetic AI focused on dissatisfaction prevention and recovery. By demonstrating how advanced ML techniques, such as cross-attention, can capture and synthesize these multimodal signals, we provide a concrete example of how theoretical models of emotion and customer dissatisfaction can be operationalized in a practical, computational framework that enables real-time intervention before dissatisfaction leads to customer churn or negative outcomes.

4.4.2 Managerial implications

In the dynamic landscape of service businesses, customers expect fast and excellent service. Based on the insights derived from this chapter, organizations can act on dissatisfaction signals in real time, improving recovery rates, optimizing agent allocation, and informing channel design decisions. Eventually, this will foster a positive and long-lasting relationship between customers and the company (Fraering & Minor, 2013). This proactive approach is particularly valuable given that dissatisfied customers are more likely to share negative WOM and switch providers, making early detection critical for protecting brand reputation and customer retention (Azemi et al., 2020; Zorn et al., 2010).

Service managers, equipped with a comprehensive understanding of customer dissatisfaction, can leverage this intelligence to oversee and improve the entirety of service interactions (Lervik Olsen et al., 2014). Unlike traditional satisfaction surveys that suffer from response biases and delayed feedback, our multimodal approach enables continuous monitoring of all customer interactions without relying on voluntary customer participation (S. Han & Anderson, 2020; K. Park et al., 2018). Utilizing natural language processing for tasks like topic modeling and automatic call routing can distill critical insights from voice interactions, facilitating preemptive action against recurrent service issues (Aksin et al., 2009; Shah et al., 2023). Linking dissatisfaction detection with behavioral data, such as customer churn probability and service agent performance metrics, enables a more strategic approach to service management by revealing the early warning signals that predict customer defection and service failures. This connection allows managers to make data-driven decisions that not

only improve immediate service interactions but also improve sustained customer engagement and optimize workforce performance.

Our dissatisfaction detection model serves as a decision support tool that can aid call center executives in identifying high-risk interactions before they escalate. This capability is particularly valuable because it enables managers to intervene immediately after the service interaction itself, rather than waiting for post-interaction surveys that many dissatisfied customers never complete. Another aspect in this context is immediate service recovery, where strategies like compensation or apologies can be triggered automatically when customers are identified as dissatisfied during the interaction. This real-time approach to service recovery represents a significant advancement over traditional methods that rely on customers to complain or complete surveys proactively (Borah et al., 2020). Research could further investigate how immediate, AI-triggered recovery strategies impact dissatisfaction mitigation (Wong, 2004; Yayla-Küllü, Tansitpong, Gnanlet, McDermott, & Durgee, 2015). Understanding the contextual effectiveness of these strategies empowers managers to tailor training, ensuring service agents are adept not only at resolving current issues but also at recognizing and responding to early dissatisfaction signals (I.-C. Lee, Lu, Fu, & Teng, 2017; Pontes & O'Brien Kelly, 2000).

A detailed analysis of real-time dissatisfaction patterns may also allow managers to allocate resources to retain high-value customers when intervention is most effective. Investigating the economic impact of immediate dissatisfaction intervention offers a strategic advantage, enabling organizations to prevent customer churn at a fraction of the cost of traditional customer acquisition (Lemmens & Gupta, 2020). This real-time approach promises not only operational efficiency but also the ability to transform potentially churning customers into loyal advocates through responsive service recovery.

Finally, our exploration into dissatisfaction detection extends to offering immediate support for service agents during customer interactions, as real-time alerts about emerging dissatisfaction can help service agents adjust their approach before the situation deteriorates (Bromuri et al., 2021; De Ruyter et al., 2001). We submit it to future research to further investigate how real-time dissatisfaction detection can offer immediate support to service agents, enhancing their ability to de-escalate tense situations and prevent customer churn during the interaction itself. Integrating systems that alert agents to early dissatisfaction signals can help organizations create a supportive environment that enables proactive problem resolution rather than reactive

damage control.

4.4.3 Limitations and suggestions for future research

While the findings provide valuable insights for distinguishing satisfied from dissatisfied customers within voice-based service interactions, the study does not come without limitations, which inform fruitful avenues for future research. First, service interactions were labeled by independent annotators, which captures displayed customer dissatisfaction rather than direct customer feedback, which is typically obtained through surveys. This introduces a layer of subjectivity, as it may produce instances where the objective coding may not be aligned with the actual customer sentiment, despite the relationship between customer satisfaction and sentiment (Y. Kim et al., 2020). Although customer-labeled data also carries inherent biases (e.g., gender bias, (Moshavi, 2006); racial bias (Hekman, Aquino, Owens, & Mitchell, 2017); and anticipated firm interaction, (Mukherjee, Burnham, & King, 2021)), we recommend future research to validate our findings among a sample of customer-labeled data. Such research could also account for the cultural nuances in customer dissatisfaction expression and interpretation, as customers may respond differently to predefined scripts and service recovery strategies (Tombs, Russell-Bennett, & Ashkanasy, 2014). Future research could also further investigate how human annotators' tendency to detect negative emotions more accurately, known as negativity bias, affects model training. For example, studies could explore whether AI performance differs when trained on datasets with more subtle or ambiguous negative instances, or when comparing human-labeled and customer-labeled data.

Second, the sample size and imbalanced nature of the customer dissatisfaction labels could have affected the performance of the custom models. The audio model, in particular, faced challenges due to the variability in the audio signal, influenced by factors such as gender, speaking styles, and accents. Comparing a model trained on a smaller dataset to a model trained on a substantially larger one presents inherent limitations. Enhancing the dataset by increasing the sample size and diversity could further improve model performance.

A different type of approach for future research involves utilizing multimodal large language models, which employ cross-attention to integrate various data types. Presently, a few promising open-access models, including Salmonn (Tang et al., 2024) and Macaw (Lyu et al., 2023), demand hardware capabilities beyond what is currently fea-

sible. Multimodal models from Meta and OpenAI, which are not free and restricted due to GDPR and privacy concerns, are unsuitable for immediate use. However, these models may become viable options for future research as they evolve to better suit current tasks.

Finally, customer dissatisfaction is not static and may fluctuate during a call (Ashtar et al., 2023). This dynamic nature of satisfaction and dissatisfaction highlights the potential for examining patterns and shifts in customer sentiment throughout the call, offering valuable insights into effective scripting, procedures, and strategies for service agents to enhance customer satisfaction. Additionally, customers may experience satisfaction with one aspect of the service and dissatisfaction with another, leading to a multi-dimensional customer satisfaction label. Exploring these fluctuations not only provides practical implications for service improvement but also emphasizes their significance for future research.

In conclusion, while our study advances the understanding of detecting customer dissatisfaction from call center service interactions, many avenues remain for future research to further explore the complexity and multifaceted nature of this domain. Additional insights may contribute to developing more effective tools for the analysis and management of customer dissatisfaction in service interactions (Lervik Olsen et al., 2014).

Chapter 5

Automated detection of firm social media response strategies: A multi-label classification study of X-based customer service interactions

This chapter is based on an article accepted as 'Waelbers, B., Henkel, A. P., & Bromuri, S. (in press). Automated detection of firm social media response strategies: A multi-label classification study of X-based customer service interactions. In *10th International Conference on Machine Learning Technologies* (pp. 310-315). IEEE.'

Abstract

A considerable share of consumer-firm interactions today unfolds online. It is therefore indispensable for firms to understand the implications of their social media responses, especially when addressing negative consumer sentiment. A first step is to identify the response strategies firms deploy. This study is the first to incorporate large language models (LLMs) alongside deep learning to extract response strategies of service firms from their social media interactions with consumers. Employing the EmoTwiCS dataset, seven strategies are extracted from 5,299 interactions on X (formerly Twitter). We approach strategy identification as a multi-label classification problem since multiple strategies can be used in a single conversation. We compare deep learning models with various types of embeddings and LLMs to extract these strategies. Our results show that while LLMs with examples perform reasonably well, a custom-trained multi-layer perceptron model using bag-of-words representations performs best. This research offers valuable insights for future studies and organizations looking to analyze the effects of response strategies on service outcomes such as customer satisfaction, emotions, and the service recovery process.

5.1 Introduction

Social media has become a key platform for consumers to engage with companies (Einwiller & Steilen, 2015). A prominent example is X (formerly Twitter) with approximately 611 million active monthly users (Duarte, 2024). With 83% of consumers expecting a firm's response on social media, and half of the consumers willing to switch brands due to poor service, companies need effective social media strategies (Bohne, Raphael, 2022; Navarro, 2023).

Yet, unlike private communication via channels, such as phone calls or emails, social media interactions are publicly available to other marketplace participants (Van Herck, Decock, & De Clerck, 2020). Given the public nature of social media posts, negative word-of-mouth (WOM) in the form of unfavorable comments and complaints directed at companies emphasizes the need for firms to respond effectively (Xun & Guo, 2017). Companies must resolve individual complaints and simultaneously deploy effective de-escalation techniques to influence consumer brand perceptions positively (Van Mulken, 2024).

Accurate identification of response strategies is essential to analyze their impact on consumer outcomes, where automation can accelerate this process for large-scale conversations. Although artificial intelligence (AI) and machine learning (ML) have been successful in extracting consumer-oriented variables, such as topics (S. Huang, Peng, Li, & Lee, 2013) and emotions (Krishnan, Elayidom, & Santhanakrishan, 2017), the automated extraction of response strategies has only received limited attention (S. N. Kim, Cavedon, & Baldwin, 2010; Oraby, Gundecha, Mahmud, Bhuiyan, & Akkiraju, 2017; Oraby et al., 2019). This lack of extensive research with more recent technologies is notable given the established importance of response strategies for customer satisfaction (Einwiller & Steilen, 2015), service recovery (Istanbulluoglu & Oz, 2023), and WOM (Xun & Guo, 2017).

In this chapter, we aim to automate the extraction of a firm's social media response strategies by framing it as a multi-label classification problem, as multiple strategies can be present within a single conversation. Our approach uses deep learning and large language models (LLMs). We utilize the EmoTwiCS dataset (Labat et al., 2024), featuring 5,299 firm-consumer interactions on X that are labeled with eight firm response strategies: 1) apology, 2) cheerfulness, 3) empathy, 4) explanation, 5) gratitude, 6) help offline, 7) request information, and 8) other.

Our study offers several contributions. First, we introduce LLMs for multi-label

classification in firms' social media response strategies, investigating their ability to categorize multiple strategies using various model prompts. By exploring LLMs alongside deep learning models, we aim to broaden the scope of methods available for analyzing such interactions, enhancing the understanding of LLMs in interpreting customer service interactions (S. N. Kim et al., 2010; Oraby et al., 2017). Second, we conduct a comparative analysis between deep learning models utilizing different embeddings and LLMs. We aim to identify their respective strengths and weaknesses in extracting response strategies (Oraby et al., 2019). Thereby, we extend insights into the most effective methods for extracting information from social media service interactions (Balaji, Annavarapu, & Bablani, 2021). Third, using a public repository of real-world X conversations enhances external validity (Labat et al., 2024). It also presents a benchmark for future models. Fourth, we contribute back to the discussions on the evolving role of AI in service management, indicating that technology delivers valuable insights that can aid businesses in improving their service interactions by not only extracting customer information but also company response strategies (Bromuri et al., 2021; Henkel, Bromuri, et al., 2020; L. Ma & Sun, 2020).

5.2 Related work

With the growing importance of social media in the domain of customer service, companies need efficient ways to handle large volumes of interactions (Dobrucali Yelkenci, Özdağoğlu, & İlter, 2023). Firms are increasingly applying AI to extract insights from service interactions (L. Ma & Sun, 2020). Research has particularly focused on customer complaints, as related interactions provide critical insights into opportunities for service improvement (Cambra-Fierro, Melero, & Sese, 2015).

AI can be used to extract valuable insights from social media messages, enabling firms to understand their customers' attitudes and issues better. For example, decision trees can be used to analyze customer complaints to uncover the root causes of service failures (Cambra-Fierro et al., 2015). Another study applied ML techniques, such as pretraining and fine-tuning, on 865,000 tweets for five customer-related tasks, highlighting ML models' scalability and versatility in processing large volumes of social media data (Hadifar, Labat, Hoste, Develder, & Demeester, 2021). Also, multi-label classification has been applied for extracting insights from customer posts (e.g., in customer reviews), aiming to extract the different customer opinions

(Deniz, Erbay, & Coşar, 2022). Similarly, sentiment and topics can be seen as a multi-label classification problem, where different topics can be related to various customer sentiments (S. Huang et al., 2013). These methodologies demonstrate the effectiveness of multi-label classification in understanding customer service dynamics. Altogether, these studies illustrate the significant focus of research on the deployment of ML and AI to analyze customer contributions on social media, primarily from the customers' perspective.

In comparison, response strategies remain relatively under-explored in the context of automation. While some studies have examined how companies respond to customer tweets, identifying strategies such as apology, information request, and empathy (Einwiller & Steilen, 2015; Istanbulluoglu & Oz, 2023), these efforts are still limited in scope. For instance, dialogue acts (e.g., greeting, answer) have been automatically recognized using ML methods like naive Bayes, linear support vector classifier, and hidden Markov models (Oraby et al., 2017). These classifications were also linked to conversation outcomes (Oraby et al., 2019). Similarly, dialogue acts in live chats have been classified using features such as bag-of-words (BoW), structural information, and utterance dependency, with BoW performing best (S. N. Kim et al., 2010). However, despite these valuable contributions, more recent technological advancements, such as LLMs, have largely been overlooked. This underscores the need for exploring the potential of LLMs to enhance the automation and effectiveness of social media response strategies.

5.3 Dataset

The current study utilizes the EmoTwiCS corpus, a comprehensive open-access dataset consisting of 9,489 Dutch customer service dialogues scraped from X (Labat et al., 2024). The messages were collected in 2020 from accounts of service firms operating in telecommunications, public transportation, and airline sectors active in Belgium. In total, there are 12,715 customer utterances and 13,067 firm utterances. The dataset is annotated with multiple layers of emotional and contextual information, including emotion labels; scores for valence, arousal, and dominance; conversation causes; and response strategies. For this study, we focus exclusively on the response strategy annotations, since other labels are not typically available in real-world settings. This results in 5,299 conversations that contain one or more labeled

response strategies out of eight possible strategies (i.e., apology, cheerfulness, empathy, explanation, gratitude, offline assistance, request for information, and other). Emojis were not included in the analysis, as they were not represented as characters in the dataset and could not be processed adequately within the models.

The Mistral-7B-Instruct model was selected to translate all conversations into English because both the underlying model and its embeddings were primarily trained on English data. Additionally, its open-source nature ensures transparency, and its size allows it to run on a high-performance laptop. The translations were reviewed by a native Dutch speaker and verified to be of high quality.

For multi-label classification problems, the data can be described by two key metrics: label cardinality and label density. On average, a conversation was labeled with two strategies, meaning that the dataset has a label cardinality of 2. The label density, which refers to the proportion of labels per data point, is 0.25. The conversation with the highest number of strategies was labeled with seven strategies.

Since deep learning requires a training sample, the dataset was divided with an 80-20 split for training and testing. All models were evaluated using the same test set to ensure fair comparison, despite LLMs not requiring a training set. Table 5.1 presents the occurrences per strategy in both the training and test sets.

Table 5.1: Response strategies occurrences

Response Strategy	Apology	Cheerfulness	Empathy	Explanation	Gratitude	Help offline	Request info.	Other
Train (%)	517 (80%)	945 (81%)	799 (79%)	2,682 (80%)	409 (80%)	1,657 (79%)	1,008 (80%)	189 (83%)
Test (%)	131 (20%)	225 (19%)	192 (21%)	656 (20%)	102 (20%)	452 (21%)	260 (20%)	40 (17%)
Total	648	1,170	911	3,338	511	2,109	1,268	229

5.4 Methodology

5.4.1 Multi-label classification

Multi-label classification is a type of ML where for every input, multiple labels can be annotated: Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector and $\mathbf{y}_i \in \{0,1\}^L$ is a binary vector indicating the presence or absence of labels, with L being the total number of labels. The objective is to predict $\mathbf{y}_i$ for each instance $\mathbf{x}_i$.

For every conversation, multiple strategies can be identified. The label set for conversation i can be represented as:

$$\mathbf{y}_i = [y_{i1}, y_{i2}, \cdots, y_{iL}]$$

Where $y_{ij} = 1$ if label j is associated with conversation i, otherwise $y_{ij} = 0$.

5.4.2 Deep learning

In this study, we implement a deep learning model to perform the multi-label classification on extracted features from the translated conversation data (see Figure 5.1). Various embedding extractions and representations (BoW, RoBERTa, and Mistral) are used to transform the textual conversation data into numerical representations, which are fed into a multi-layer perceptron (MLP) for training and prediction (see section 5.4.2). The outcomes of the models are represented in one-hot encoding format, where each position corresponds to a single strategy.

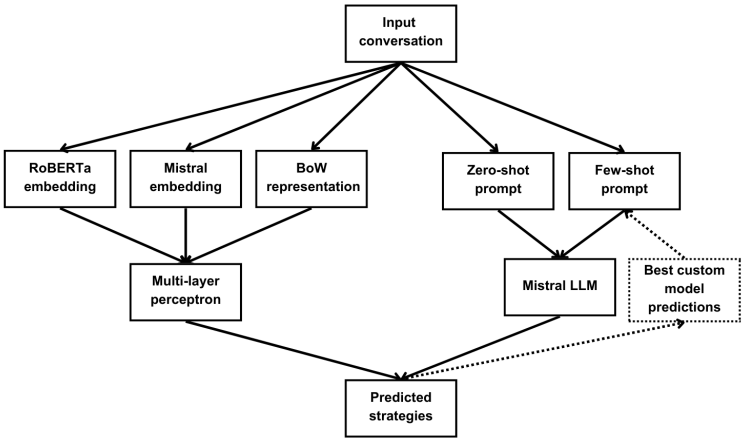

Figure 5.1: Overview methods

BoW representations

BoW representations represent textual data in relation to their frequencies, regardless of grammar and word order, into numerical representations. For extracting the BoW features, the *CountVectorizer* from Python's *scikit-learn* library was used. Here, the texts of the conversations are transformed into numerical representations based on the frequency of the words. It results in a sparse matrix where each row represents a conversation and each column represents a word from the vocabulary.

RoBERTa embeddings

A second type of embedding was extracted using the RoBERTa model (Y. Liu et al., 2019). RoBERTa is an optimized version of BERT that uses self-attention mechanisms to process and encode the relationships between words in a sentence, capturing bidirectional context to provide a nuanced representation of the text.

Mistral embeddings

The third embedding was obtained from the Mistral-7B-Instruct-v0.3 model (Jiang et al., 2023). This model is an open-source LLM that can run on a GPU-equipped computer, making it a transparent and accessible option.

Model architecture

The deep learning MLP model consists of two linear layers. Xavier uniform initialization was utilized for the weights. The first layer employed a ReLU activation function, with a dropout rate of 0.5 to prevent overfitting. Finally, a separate sigmoid function was applied. The model has a hidden dimension of size 50 and an output layer of size 8, which denotes the eight strategies in one-hot encoding. An Adam optimizer was used with a learning rate of 0.001. A learning rate scheduler was implemented with a step size of 10 and a decay factor of 0.7. For training, the binary cross-entropy loss was implemented. The training was conducted for 30 epochs, with a batch size of 64. To prevent exploding gradients, gradient clipping was applied.

5.4.3 Large language model

Model

In addition to training a deep learning model, we investigate whether a fully pre-trained LLM can perform multi-label classification of a service firm's strategies from X customer service conversations. Three different open-access LLMs were examined: Mistral-7B-Instruct (Jiang et al., 2023), Google's Gemma-7B-Instruct (Gemma Team et al., 2024), and Phi-Instruct-mini (Microsoft, 2024). All models were evaluated, but we only display the results of the best model, Mistral-7B-Instruct, for brevity (see Figure 5.1).

Zero-shot

Zero-shot prompting refers to giving the model a prompt that does not include any examples or demonstrations. However, some context is provided for each strategy to help the model better understand them. This results in the following prompt:
The following text is an interaction between a customer and a service agent. Please choose the most suitable strategies that the service agent uses. There can be multiple strategies, but just give the most important ones that are explicitly in there (no implied or implicit strategies).
Answer with the strategies, do not give a sentence or an explanation, nor the strategies that were not used.
Please base your answer on the following conversation:

Few-shot

Few-shot prompting provides the model with examples to guide its performance. This study incorporated an English example of each strategy in the zero-shot prompt, followed by one combining multiple strategies.

Few-shot with BoW prediction

Next to context and examples, a prompt can include predictions from another model. Here, the few-shot LLM prompt is supplemented with predictions from the best custom-trained model, enabling it to incorporate this information into its reasoning for a prediction.

5.4.4 Evaluation metrics

Hamming score, Jaccard index, and average exact match are used as evaluation metrics. F1-scores are computed to show the results per strategy.

Hamming score calculates the fraction of correctly predicted labels.

$$\text{Hamming score} = \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{j=1}^{L} I(y_i^j = \hat{y}_i^j) \tag{5.1}$$

where N is the number of samples, L is the number of labels, y_i^j is the true label for the i-th sample and j-th label, $\hat{y}_i^j$ is the predicted label for the i-th sample and j-th label, and $I(y_i^j = \hat{y}_i^j)$ is an indicator function that equals 1 if $y_i^j = \hat{y}_i^j$ and 0 otherwise.

Jaccard index measures the proportion of shared labels (intersection) between predicted and true label sets relative to their total unique labels (union) in both sets.

$$\text{Jaccard index} = \frac{|A \cap B|}{|A \cup B|} \tag{5.2}$$

where A and B are two sets, $|A \cap B|$ is the size of the intersection of sets A and B, and $|A \cup B|$ is the size of their union.

Average exact match measures the proportion of instances where the predicted labels match the true labels exactly, averaged across all instances.

$$\text{Average exact match} = \frac{1}{N} \sum_{i=1}^{N} I(\hat{y}_i = y_i) \tag{5.3}$$

where N is the number of samples, $\hat{y}_i$ is the predicted set of labels for the *i*-th sample and y_i is the true set of labels for the *i*-th sample.

5.4.5 Model comparison

A comparative analysis of each model pair assessed differences in predictions. The Hamming score, Jaccard index, and average exact match were calculated across 10 folds to enable direct performance comparisons. A paired t-test then determined whether the models' predictions differed significantly on the test set.

5.5 Results

For the deep learning models, the BoW representations approach shows the best performance. The RoBERTa and Mistral embedding models fail to identify all strategies, with gratitude not being recognized by either model, as depicted in Table 5.2 by the 0.0000 cells. Therefore, these types of embeddings are not well-suited for recognizing response strategies accurately, since it is vital to perform well on every single strategy.

Adding more information to the LLM prompt improves its performance. The zero-shot model has a Hamming score of 0.66, the few-shot model improves to 0.70, and the combined model reaches 0.74. Similar results are seen with the average exact match, where the zero-shot, few-shot, and combined model achieves 0.0189, 0.0481, and 0.0972, respectively. However, the Jaccard index shows a different pattern: the combined model has the highest score at 0.4389, while the zero-shot and few-shot models have scores of 0.3961 and 0.3894, respectively. The weighted F1-scores are similar across all LLM models.

Overall, the custom-trained model with BoW representations outperforms all models. This is likely because BoW effectively captures specific keywords of each strategy, such as “sorry” and “apologize” for apologetic response, or “thanks” and “happy” for gratitude. These words are critical in distinguishing conversational strategies, which more complex models may overlook. While LLMs are powerful, they might not always prioritize such explicit lexical patterns, making BoW a more suitable choice in small, structured X conversations.

Table 5.2: Performance metrics comparison

Metric		BoW	RoBERTa	Mistral	Zero-shot	Few-shot	BoW & Few-shot
Hamming score		0.8783*	0.8600	0.7607	0.6597	0.7031	0.7407
Jaccard index		0.3764*	0.3415	0.0594	0.0189	0.0481	0.0972
Average exact match		0.6322*	0.5872	0.3946	0.3961	0.3894	0.4389
F1	Micro	0.7185	0.6520	0.5267	0.5336	0.5292	0.5662
	Macro	0.5554	0.3400	0.2096	0.4503	0.4527	0.4645
	Weighted	0.6915	0.5666	0.4046	0.5991	0.5835	0.6033
	Apology	0.7926	0.0725	0.0645	0.5111	0.6416	0.4929
	Cheerfulness	0.5944	0.6006	0.2674	0.4144	0.4259	0.5081
	Empathy	0.4421	0.0104	0.000	0.3197	0.2924	0.3318
	Explanation	0.8288	0.8287	0.7646	0.8267	0.7795	0.7875
	Gratitude	0.3910	0.0000	0.0000	0.2651	0.2070	0.2676
	Help offline	0.8605	0.8423	0.5806	0.7023	0.6586	0.7106
	Request info.	0.4851	0.3659	0.0000	0.4641	0.5000	0.4960
	Other	0.0488	0.0000	0.0000	0.0988	0.1167	0.1215

Note: * indicates a statistically significant difference of the best performing model vs. all other models at $p<0.001$. F1-scores are displayed for illustration.

Table 5.3: Model comparison for Hamming score, Jaccard index, and average exact match

	RoBERTa	**Mistral**	**Zero-shot**	**Few-shot**	**BoW & Few-shot**
BoW	0.0183^{***} 0.0349^{***} 0.0450^{***}	0.1176^{***} 0.3170^{***} 0.2376^{***}	0.2186^{***} 0.3575^{***} 0.2361^{***}	0.1752^{***} 0.3283^{***} 0.2428^{***}	0.1376^{***} 0.2792^{***} 0.1933^{***}
RoBERTa		0.0994^{***} 0.2821^{***} 0.1926^{***}	0.2003^{*} 0.3226^{***} 0.1911^{***}	0.1569^{***} 0.2934^{***} 0.1978^{***}	0.1193^{***} 0.2443^{***} 0.1483^{***}
Mistral			0.1010^{***} 0.0405 -0.0015^{**}	0.0576^{***} 0.0113^{***} 0.0052	0.0200^{***} -0.0378^{***} -0.0443^{**}
Zero-shot				-0.0434^{***} -0.0292^{***} 0.0067^{***}	-0.081^{***} -0.0783^{***} -0.0428^{***}
Few-shot					-0.0376^{***} -0.0491^{***} -0.0495^{***}

Note: Values indicate the delta (row vs. column) of the evaluation score
with *** p$<$0.001, ** p$<$0.01, * p$<$0.05.
Per cell, first row: Hamming score, second row: Jaccard index, third row: average exact match.

When performing statistical tests for model comparisons, all models are significantly different when using the Hamming score (Table 5.3). For the Jaccard index, only the Mistral embeddings model and the zero-shot model show no significant difference. The overall Jaccard index for these models is 0.3946 and 0.3961, respectively (Table 5.2), showing that both models perform equally in terms of Jaccard index. Regarding the average exact match, only the Mistral embeddings model and the few-shot model are not significantly different from each other (Table 5.3). Table 5.2 shows that the average exact match for the Mistral embeddings model and the zero-shot model have values of 0.0594 and 0.0481, respectively. This proximity in scores suggests comparable performance in achieving exact matches between predicted and true label sets.

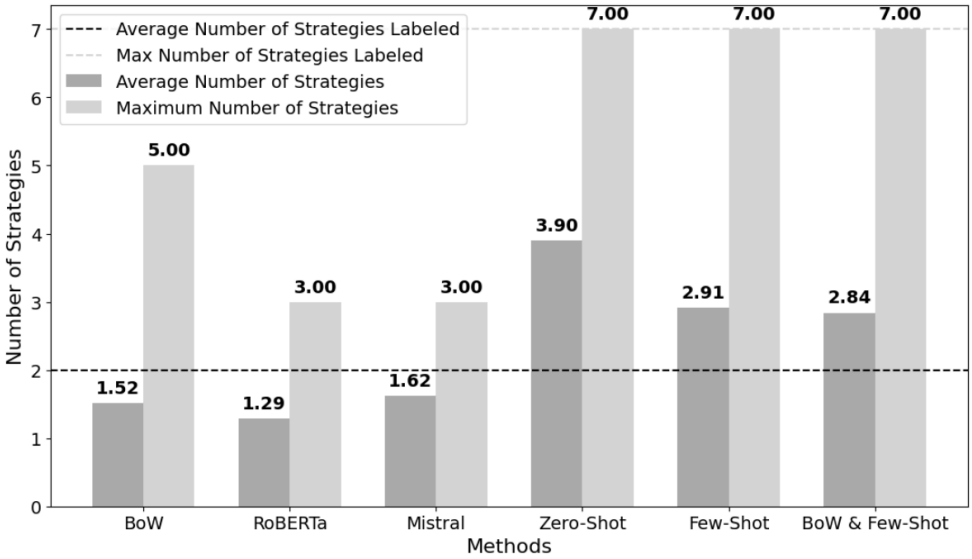

Figure 5.2: Average and maximum number of strategies per method

When examining the number of strategies predicted, substantial differences emerge, as can be seen in Figure 5.2. The deep learning models generate notably fewer predictions (below two strategies on average) per conversation compared to the dataset's labels (2 strategies on average) and the LLMs (above two strategies on average). Similar results can be found for the maximum number of strategies predicted: 7 for the dataset labels and for the LLM, 5 for the BoW representations model, and 3 for the RoBERTa and Mistral embeddings models.

5.6 Conclusion and future works

This chapter extracts social media response strategies of a service firm's interactions with its customers on X as labeled in the EmoTwiCS dataset (Labat et al., 2024). Training a deep learning model with BoW representations yields superior performance compared to alternative models. Additionally, while LLMs show reasonable performance, particularly when supported with additional examples or predictions of the best-performing model, they do not exceed the best model's performance. Performance differences may arise from the different features, as BoW uses word frequency, while RoBERTa and Mistral capture context and semantics. Though effective at understanding nuanced text, LLMs may struggle to identify distinct strategies, potentially overlooking crucial keywords. This demonstrates that the largest and most sophisticated model does not always deliver the best performance. Therefore, it is crucial to consider not only LLMs but also to explore the potential of smaller, simpler models.

This chapter offers at least four contributions. First, we introduce LLMs as a novel approach for multi-label classification, specifically in firms' social response strategies (Balaji et al., 2021). We demonstrate that while LLMs do not outperform custom-trained models, they still deliver acceptable results, as they eliminate the need for separate training and extensive data labeling. Second, we provide a comparative analysis of ML models, showing that BoW representations in a custom-trained model demonstrate superior performance. Hereby, we show that the largest model does not always outperform simpler models, thereby extending insights into the respective strengths and weaknesses of different models for multi-label classification, especially in recognizing firms' social response strategies (Oraby et al., 2017). Third, the results of this study act as a benchmark for further studies on recognizing strategies from customer conversations on X (Labat et al., 2024). Finally, we contribute to service management and marketing by extending our understanding of AI techniques for managing service interactions (Bromuri et al., 2021; Henkel, Bromuri, et al., 2020).

From a service practice perspective, understanding response strategy effectiveness is essential for enhancing service protocols (L. Ma & Sun, 2020). The automated strategy extraction offers service firms a novel tool to analyze and improve their customer service interactions (Balaji et al., 2021; L. Ma & Sun, 2020). Such information is currently unavailable without manual labeling. Once recognized, these strategies can be studied individually, focusing on effective deployment in the interaction context

(Einwiller & Steilen, 2015). Implementing this information into their customer relationship management systems enables real-time identification of strategies. Firms can implement this to analyze strategy effectiveness across different customer segments and contexts (Balaji et al., 2021), and to inform data-driven decisions about which approaches best address specific customer concerns. The strategies can also be mapped on key performance indicators, such as satisfaction and emotions (Krishnan et al., 2017), service recovery (Istanbulluoglu & Oz, 2023), and can be used to optimize firm outcomes, create targeted training programs, and revise service scripts (L. Ma & Sun, 2020).

Future research could build on our work by examining larger models or by fine-tuning pretrained models, as their increased size may offer improved capabilities, potentially leading to better results. Additionally, examining other datasets might help determine whether the findings are generalizable across different contexts and domains. Similarly, additional measures from this dataset (e.g., customer emotions, valence-arousal levels, or customer intent) or alternative measures (e.g., social media usage patterns or discussion topics) could be explored to enhance the identification of response strategies (Cocarascu & Toni, 2018; Labat et al., 2024; Waelbers et al., 2022).

While we demonstrate that strategies can be derived from brief X customer service interactions, future research should explore longer chat or phone conversations, which may employ different strategies. Firms often encourage private issue resolution, which could reveal additional insights into customer emotions and problem resolution rates (Van Herck et al., 2020). Public interactions may focus on brand image and efficiency (Guo, Fan, & Zhang, 2020), whereas private interactions allow for personalized engagement (Wolf & Steul-Fisher, 2023). Studies on omni-channel service (Goffin, 1999) suggest that strategy effectiveness varies by channel, influencing customer satisfaction and firm reputation. Understanding these differences would enhance research on customer service strategies, not only for human service agents, but also for service bots (Castelo, Boegershausen, Hildebrand, & Henkel, 2023).

Chapter 6

Beyond traditional quality monitoring in customer service interactions: A comparative analysis of human evaluators and large language models

This chapter is based on an article published as 'Waelbers, B., Henkel, A. P., & Bromuri, S. (2026). Beyond traditional quality monitoring in call centers: A comparative analysis of human evaluators and large language models. In: Li, S. (eds) Information Management. ICIM 2025. Communications in Computer and Information Science, Vol. 2540. (pp.1-10). Springer, Cham. https://doi.org/10.1007/978-3-031-99353-4_27'

Abstract

In call centers, thousands of calls occur daily, necessitating robust call quality monitoring processes. This study explores the potential of large language models (LLMs) to improve these processes by supporting the human evaluator. We investigate how LLMs can help identify cases where human evaluators may have made errors in quality monitoring forms, focusing on their effectiveness in flagging potential mistakes. Our research employs a three-step approach: comparing the model's assessments with primary human evaluator decisions, validating discrepancies through blind secondary human reviews, and analyzing patterns in cases where the LLM flags potential human errors. This methodology aims to uncover both the advantages and limitations of integrating LLMs into quality monitoring processes. Hereby, we try to understand when the model shows human oversight and when it underperforms. Our findings provide insights into the advantages and disadvantages of LLMs as supportive tools in call quality monitoring, emphasizing the potential of human-AI collaboration in quality monitoring, leading to a less variable and more reliable process.

6.1 Introduction

Voice-based customer service interactions remain a cornerstone of customer-firm interactions: 54% of US consumers still mention phone calls as their preferred means of communication with companies, highlighting the importance of analog communication for the service industry (Bohne, 2024). To meet this demand, a workforce of 2.86 million contact center employees is active in the US alone, underscoring the massive scale at which service interactions occur daily (Bohne, Raphael, 2024). Within this context, maintaining consistent service quality is crucial yet increasingly challenging for companies, as every single interaction has the potential to significantly impact customer satisfaction and loyalty (Bloemer, de Ruyter, & Wetzels, 1999; Caruana, 2002).

The process of call quality monitoring presents a unique set of challenges for firms, where feedback mechanisms must address multiple dimensions of evaluation (Aksin et al., 2009). Call center interactions are typically assessed through standardized quality monitoring forms, which break down the evaluation into distinct categories such as technical knowledge, following procedures, and communication skills (White & Roos, 2005). With thousands of customer service interactions occurring daily in a single call center, evaluating and providing meaningful feedback on these interactions becomes an increasingly complex task (Tovar, 2021).

Current approaches mainly rely on human evaluators completing feedback forms. This process faces several limitations. First, human evaluators have limited time, resulting in only a small portion of calls being reviewed (Rivera, Qiu, Kumar, & Petrucci, 2021). As a result, call center managers typically assess a fraction of total interactions, potentially overlooking important patterns or borderline cases that could inform more targeted training. Second, this approach also leads to resource inefficiency, where evaluators spend time reviewing routine interactions that meet all criteria, while critical cases may go undetected. Since call center agents generally receive extensive training to maintain service quality, identifying the relatively rare instances where performance falls below expectations is both valuable and difficult (Elnaga & Imran, 2013). Third, human evaluators may interpret evaluation criteria differently, especially in borderline cases. Their assessments can also be affected by personal relationships with agents, fatigue, and fluctuating levels of attention during evaluation sessions (Breuer, Nieken, & Sliwka, 2013). Taken together, these limitations in traditional evaluation processes point to a need for more systematic methods of

identifying critical cases and patterns across interactions.

In this context, the emergence of Large Language Models (LLMs) in the domain of artificial intelligence (AI) presents an opportunity to systematically identify patterns and critical cases that human evaluators might overlook. The ability of LLMs to process vast amounts of text while recognizing subtle patterns and variations makes them particularly suitable for augmenting human evaluation processes (Teubner, Flath, Weinhardt, van der Aalst, & Hinz, 2023). Rather than replacing human evaluators, LLMs could help address the current constraints by systematically flagging potential boundary cases and identifying patterns that might be missed in traditional sampling approaches. This is especially valuable in high-performing environments where critical cases are rare but significant for maintaining and improving service quality.

This study aims to understand the potential role of LLMs in identifying critical cases in call center conversations by examining the differences between human and LLM evaluations. We aim to quantify the differences between LLM and human assessments in completing quality monitoring forms, particularly focusing on critical cases. To validate these differences, we will conduct blind secondary human reviews of flagged cases. Additionally, we will analyze patterns in boundary cases where the model's and human evaluations diverge. Through this approach, we seek to identify the strengths and limitations of LLM-based evaluation systems in quality monitoring contexts.

This study contributes to the field of language model evaluation in several ways. First, we introduce the use of LLMs in the field of call quality monitoring, where previous models relied on big data and deep learning technologies (A. Ahmed et al., 2024; Karakus & Aydin, 2016). We also provide a novel comparative analysis of LLMs and human assessments in critical cases, quantifying the extent of their differences. Second, we introduce a methodology for validating these differences through secondary blind human reviews, enhancing the robustness of evaluation techniques. Third, our research identifies and categorizes patterns of disagreement between the model's predictions and human evaluations, offering novel insights into the nature of these discrepancies (Y. Park, 2011). These contributions advance our understanding of the model's performance in complex evaluation tasks, highlight potential biases in both machine and human assessments, and aim to create a more reliable quality monitoring process by integrating the strengths of LLMs with human expertise (Celiktutan, Cadario, & Morewedge, 2024). Finally, we contribute to the broader understanding of

AI applications in feedback and quality monitoring processes (A. Ahmed et al., 2024; Laskar et al., 2023; Rivera et al., 2021).

6.2 Background and related work

The measurement of call center call quality is essential for the monitoring, feedback, training, and evaluation of service agents (Oztaysi, Onar, Kahraman, & Gok, 2020). This background section reviews the literature on the implementation of various technologies in the call quality monitoring process. As call centers generate larger volumes of data, researchers have developed automated methods for performance prediction and evaluation. One approach utilized a naïve Bayes classifier to forecast agent performance based on factors such as logged hours, talk time, and completed records (Valle, Varas, & Ruz, 2012). Building on this, the value of big data analytics combined with similarity measurements for performance evaluation offers a more comprehensive view of agent performance beyond individual call assessments (Karakus & Aydin, 2016). Another study combined quality monitoring forms with natural language processing to identify conversations most relevant for coaching or feedback (Laskar et al., 2023).

The application of deep learning techniques has been suggested to improve the call quality monitoring process. Two studies implemented a call ranking module using direct question answering and a maximum-entropy classifier to categorize calls as “good” or “bad” (Saon, Ramabhadran, & Zweig, 2006; Zweig et al., 2006). Another study employed deep neural networks and multi-classification techniques to detect subjective calls, aiming to provide more objective evaluations (A. Ahmed et al., 2024). Further, neural networks were applied to predict customer service satisfaction scores, introducing a more nuanced evaluation scale from “not met” to “far exceeded” (Paprzycki, Abraham, Guo, & Mukkamala, 2004). Another study explored text mining techniques on transcribed conversations, followed by support vector machines and logistic regression for call classification. Notably, this research addressed the issue of imbalanced data through cost-sensitive classification, a common challenge in call center environments (Y. Park, 2011). Recognizing the multiple dimensions of call quality, both quantitative and internal performance indicators can be integrated to predict evaluation scores using a backpropagation neural network, considering various aspects of agent performance (Hsu, Chen, Chan, & Chang, 2016).

Despite these advancements, some challenges remain. As service agents are highly trained in providing consistent service, data can be imbalanced (Y. Park, 2011). Deep learning techniques are known to have issues with highly imbalanced data; this study therefore suggests LLMs to complete call quality monitoring forms (Johnson & Khoshgoftaar, 2019). Furthermore, previous studies focus on an overall conversation score (ranging from good to bad, or a score on a 100-point scale). To gain a deeper understanding of the actual interactions, we will examine the individual call quality monitoring questions that are important for the interaction.

6.3 Methodology

6.3.1 Data

A global service company with a large call center division provided access to 244 telephone conversations, alongside manual call quality monitoring evaluations. All conversations were transcribed and anonymized, deploying the company version of OpenAI's Whisper-large model (Radford et al., 2023). In total, the conversations involved 166 service agents and 24 evaluators. The call quality monitoring forms are comprised of binary questions, where a score of 1 represents that the service agent met the required standards, and a score of 0 indicates that they did not. The call quality monitoring checklist contains questions about the performance of the service agent in the corresponding conversation (see Table 6.1). Questions about internal company processes were not considered in this study.

In Table 6.1, the *[percentage]* denotes the percentage of times a question was scored with a 1. The data shows that the performance is generally high, with all questions scoring above 90%. This reflects the high level of training and experience of service agents in dealing with customer service requests (Elnaga & Imran, 2013). Notably, some questions achieved a perfect performance rate of 100%, these questions will not be incorporated in the analysis.

6.3.2 Large language model: GPT-3.5-Turbo

As a model, the company's local OpenAI GPT-3.5-Turbo model was employed (see Figure 6.1 for the full overview) (Microsoft, 2024). The model is used in an offline company setting, ensuring that no personal details about customers and service

Table 6.1: Questions and percentage of 1's

Question Type	Percentage of 1's
Build rapport	98.0%
Addressed customer correctly & title	91.8%
Closed conversation correctly	90.6%
Used appropriate language	100%
Used appropriate tone	99.6%
Showed confidence and efficiency	100%
Took ownership	99.6%
Pursued solution	98.0%
Performed security verification	99.6%
Asked insightful questions	99.1%
Followed communication protocol	96.4%
Offered accurate information	93.7%

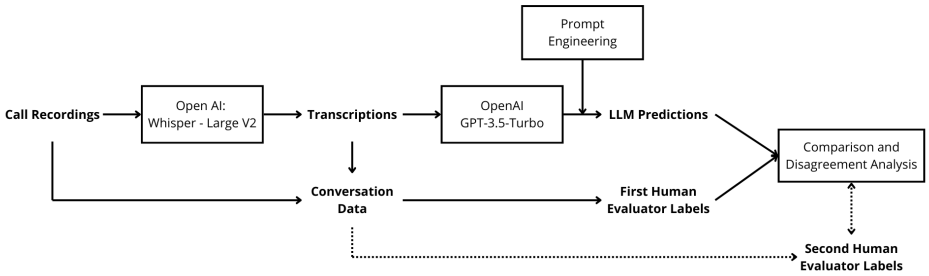

Figure 6.1: Overview of the modeling and labeling approach.

agents are transmitted. To enable the model to complete these forms effectively, we utilize prompt engineering. Prompt engineering implies adjusting the input provided to the model to optimize its outputs (Marvin, Hellen, Jjingo, & Nakatumba-Nabende, 2024). The prompts should include all information the LLM model requires to answer the questions. Adjusting the prompt is an iterative process, repeatedly testing the outcome of the model, based on the prompt, compared to a predefined baseline. For this study, an exploratory prompt engineering process was performed on a training set of 250 calls with similar questions to the themes in Table 6.1. After the reiterative process of writing these prompts, the prompts were tweaked to the current dataset.

6.3.3 Second human evaluator

Acknowledging the potential for human errors due to constraints such as time limitations, lapses in attention, and biases, a protocol for reviewing the labels was implemented when discrepancies arise between human and LLM labels. Here, a second human evaluator, blindly and without a time limit, assigned an additional label to the conversations, which was then seen as the ground truth. In total, six different evaluators labeled the disagreement cases. This iterative updating process ensures the continuous refinement of the labeled dataset, enhancing its accuracy and reliability for subsequent analyses and model training in call quality monitoring tasks.

6.4 Results

6.4.1 LLM versus human evaluator comparison

The comparison with the first human evaluator is shown in Table 6.2. It indicates a high level of agreement for the “1” class, whereas the agreement for the “0” class is significantly lower, with an F1-score of 0.033.

Table 6.2: Confusion matrix of the comparison between the labels of the LLM and the first human evaluator.

	LLM Label	
First Human Evaluator Label	1	0
1	2,295	89
0	53	3

6.4.2 Second human evaluator

A second evaluator performed a blind re-evaluation of the cases in which the first human evaluator and the LLM disagreed. The confusion matrix in Table 6.3 presents the distribution of labels between the initial human evaluator labels and the second human evaluator labels. Out of those 142 cases, the first and second human evaluators agreed on 85 instances (59.9%). Disagreement occurred in 57 cases (40.1%).

Table 6.3: Confusion matrix of the comparison between the labels of the first and the second human evaluator in cases of disagreement between the LLM and the first evaluator.

	Second Human Evaluator Label	
First Human Evaluator Label	1	0
1	53	21
0	36	32

6.4.3 LLM versus second human evaluator

The labels were updated based on the second human evaluator as the ground truth. The overall comparison is presented in Table 6.4. Out of a total of 2,440 instances, the LLM and human evaluators agreed on 2,355 cases (96.52%), with 2331 instances labeled as “1” and 24 as “0”. Disagreements occurred in 85 cases (3.48%), where the LLM predicted “0” for 53 instances that humans labeled as “1”, and conversely, predicted “1” for 32 instances that humans labeled as “0”.

Table 6.4: Confusion matrix of the comparison between the labels of the LLM and the second human evaluator.

	LLM Label	
Second Human Evaluator Label	1	0
1	2,331	53
0	32	24

The confusion matrix shows a higher degree of agreement between the LLM and the second human evaluator, with an F1-score of 0.98 for the “1” class and an F1-score of 0.361 on the “0” class, being significantly better than the F1-score of 0.033 between the LLM and the first evaluator. Overall, these results indicate that while the LLM can recognize positive cases, it faces challenges in accurately identifying negative ones. Moreover, the results suggest that the model tends to be stricter in its negative predictions compared to human evaluators. In cases of disagreement, the LLM more frequently assigned a “0” whereas humans had assigned “1”, with disagreement on 53 and 32 questions, respectively.

6.4.4 Disagreement analysis

A disagreement analysis was performed to explore the differences between the LLM and the human predictions. Although not all differences have an explainable origin, several patterns emerged. First, there were transcription problems, where transcripts contained duplicated sentences, problems with short words, and inaccuracies with names. Misspelled names particularly affected the LLM's outcomes, as slightly different names result in negative agent evaluation because of incorrect customer addresses. Another problem with the names was that the model confused the different names within the conversation, mixing up service agents', customers', and even street names in addresses (e.g., Hamilton Street, Kennedy Avenue). Then, for questions "took ownership" and "pursued solution", the LLM tended to be stricter than the human. This was especially the case when the service agent suggested to wait, or transferred the customer to another department. While the LLM considered this as not taking responsibility for the conversation, nor exploring all options for resolving the issue or answering the customer's question, a human evaluator considered that as correctly following protocol. Finally, differences arose on questions such as "used appropriate tone". Here, the human evaluator scored a "0" due to issues in the way the service agent spoke (e.g., long silences, unfriendly tone, etc.). The LLM only had access to the transcripts, so the model did not pick up these issues.

6.5 Discussion

In this chapter, we investigate whether the LLM can improve the process of call quality monitoring. When the LLM is compared with the first human evaluator, the LLM appears to perform poorly. Then, the secondary human re-evaluates the disagreements between the LLM and the original human evaluator. Here, the LLM is found to be correct in 40% of the disagreement cases, leading to a significant improvement in the LLM's overall performance, especially on the "0" class. It also underscores the model's potential to identify situations where human evaluators may make errors. This suggests that the LLM's role could be particularly beneficial in acting as a second opinion or safety net in the evaluation process, helping to reduce errors that may arise from human oversight. This capability is particularly beneficial in large-scale quality monitoring systems, where human evaluators might overlook subtle issues or experience decision fatigue (Zanzotto, 2019).

The cases of disagreement between the LLM and human evaluators revealed several patterns. First, the model's inability to incorporate audio-based information plays a significant role in these discrepancies, as humans do use these audio cues. Future models may benefit from incorporating multimodal inputs that include both audio and text data (S. Wu, Fei, Qu, Ji, & Chua, 2024). Second, the LLM is stricter in assigning "0" labels than human evaluators for certain questions, such as "taking ownership", indicating that the service agent could have done more. This stricter assessment raises questions about whether the model is identifying areas for improvement in service agent performance or simply applying overly strict standards.

The evaluation of call center conversations is subject to several data limitations. One notable limitation is the prevalence of questions with over 99% "1" labels, suggesting that the service agent's performance is close to optimal. This extreme skew in the data reduces variability and makes it difficult to evaluate the model's performance across a diverse range of scenarios. It also complicates the identification of these specific, rare cases of suboptimal performance. Additionally, the current binary labeling system limits the granularity of the evaluations. Implementing a more nuanced labeling scale, such as 1 to 5, would provide a richer framework for capturing degrees of compliance and quality. For instance, rather than classifying a performance with "one small mistake" as either fully acceptable (1) or unacceptable (0), a scale could allow for labeling a "3" for satisfactory performance or a "4" for good performance with minor issues. This approach allows evaluators to express a range of quality, reducing strict categorization and providing clearer insights for improvement, ultimately enhancing quality monitoring.

Future research could examine the reasoning behind scoring decisions by both human evaluators and the LLM, using qualitative analysis to identify discrepancies or biases. This aligns with research showing that clear explanations can reveal hidden biases in algorithmic decisions, allowing users to better understand and correct potential biases in both the LLM output and their own judgments (Celiktutan et al., 2024). It would also be valuable to explore what happens if the disagreements between the LLM and human evaluators are re-evaluated by another LLM instead of a second human. Such an approach could reveal the consistency and robustness of model reasoning across different models or prompting strategies, offering insights into how LLM agreement correlates with human judgment. Furthermore, the explanations provided by the LLM could increase the transparency and explainability. Future research may also extend LLM-based monitoring to other relevant aspects, such as

the affective content of a service interaction (Henkel, Bromuri, et al., 2020; Waelbers et al., 2022), or service agent stress and well-being work (Bromuri et al., 2021).

6.6 Conclusion

This chapter shows that the call quality monitoring process can benefit from the combination of AI models and human evaluators. LLMs show to be correctly predicting 40% of the disagreement cases between LLM and human evaluator, when checked by a second human evaluator. The research contributes to understanding human-AI collaboration, showing how AI can augment rather than replace human capabilities. However, implementing such systems requires addressing biases in current processes and algorithms (Celiktutan et al., 2024). Future work should focus on refining AI-assisted monitoring systems across various contexts, as we show that this collaboration has the potential to improve quality assurance.

Chapter 7

Curiosity-driven BDI agents for aggregated knowledge extraction with applications in customer service

---

This chapter is based on a manuscript currently under preparation for submission at *The 41st ACM/SIGAPP Symposium on Applied Computing* in collaboration with Dr. Alexander P. Henkel, Dr. Jesse Heyninck, and Prof. Dr. Stefano Bromuri.

Abstract

Ontology generation is an important task in the field of knowledge extraction, particularly in knowledge-intensive domains such as organizations, where large volumes of unstructured textual data can hinder effective decision-making and insight generation. This process traditionally relies on manual expert annotations and classifications, which can be time-consuming, especially when documents grow in volume and complexity. Recently, large language models have been suggested due to their abilities to understand and process natural language text. However, these methods are often fixed, passive reasoning approaches that require access to the whole document set. In this study, a belief-desire-intention (BDI) agent is proposed as a curiosity-driven, novelty exploration methodology for ontology generation. The BDI agent relies on large language models to dynamically generate questions that guide the retrieval of relevant documents, based on its beliefs, desires, and intentions. The Text2KGBench dataset is utilized to compare our BDI model against existing ontology generation models, namely the Hearst pattern extraction and the Rebel methods. Our model consistently outperforms both baselines across continuous and graph-based evaluation metrics, demonstrating its effectiveness in capturing structured knowledge. Our work contributes to the ongoing development of adaptive systems for ontology generation, offering a promising direction for more scalable, goal-directed knowledge extraction. Beyond its technical contribution, this approach demonstrates how knowledge-intensive organizations, including service firms, can dynamically organize large collections of internal resources (e.g., FAQs, troubleshooting guides, policy documents) into actionable knowledge structures. By connecting customer-facing issues with these organizational knowledge maps, the method supports more effective service delivery, decision-making, and organizational learning.

.

7.1 Introduction

The formalization of domain knowledge into structured formats is an essential objective in knowledge representation. An ontology, a formal description that defines the concepts and relationships relevant to a particular domain of knowledge, serves as a key tool in achieving this objective (Gruber, 1995; Rao & Georgeff, 1991). Although ontologies are essential in knowledge representations, building and maintaining them is a challenging process. Traditionally, methods have relied on manual construction by human experts (Tudorache, 2019). However, this approach entails several issues, including inconsistencies and scalability limitations (Seddon & Srinivasan, 2014). Moreover, manual ontology development is resource-intensive and requires significant time and expertise (Tudorache, 2019). Unstructured data is rapidly growing, making these manual ontology-building methods unsuitable for handling large-scale, dynamic datasets (Seddon & Srinivasan, 2014).

Whereas earlier chapters focus on signals embedded in real-time customer interactions, service organizations also rely on extensive background documentation and records, including training manuals, quality monitoring forms, product knowledge bases, regulatory guidelines, as well as textual artifacts generated by customers, such as reviews, complaints, or feedback (Alaimo & Kallinikos, 2021). Making sense of this diverse and often unstructured information at scale is critical for service agents and managers (Osman, Mohd Noah, & Saad, 2022). By aggregating these documents into a structured layer, our approach demonstrates how curiosity-driven ontology generation can dynamically organize organizational knowledge, linking customer-facing exchanges with internal resources. This aggregated perspective enables a multi-layered view of service interactions, supporting improved decision-making, knowledge management, and the continuous refinement of organizational practices.

Before one can directly apply this curiosity-driven ontology generation to organizational knowledge, we first validate the method on benchmark datasets to assess its performance rigorously. This step ensures that the approach is reliable before it is applied to real-world organizational documents, where it could support knowledge structuring and link operational insights with customer-facing interactions.

Artificial intelligence (AI), and in particular large language models (LLMs), have emerged as a promising tool to support the ontology construction process. Due to their advanced language capabilities, LLMs can assist with various steps of auto-

mated ontology generation from unstructured data, as LLMs have been explored as tools to interpret complex and unstructured data (X. Liu, Sun, Lei, & Zhu, 2024), to extract relevant knowledge from text (D. Xu et al., 2024), and to understand context (An et al., 2024). These capabilities suggest that LLMs could be valuable for the different tasks or the whole process of ontology development (Babaei Giglou, D'Souza, & Auer, 2023).

Although recent advances in LLMs have opened new possibilities for knowledge extraction from unstructured data, current systems largely remain passive, retrieving information without active reasoning or dynamic exploration of new knowledge. To address these limitations, we propose a novel model inspired by cognitive agent architectures, specifically the Belief-Desire-Intention (BDI) framework (Kashima, McKintyre, & Clifford, 1998; Mele, 1989). The BDI model is a widely studied framework in the domain of agent architectures, where it provides a natural foundation for reasoning about an agent's current knowledge (beliefs), objectives (desires), and decision-making strategies (intentions). In the current context of ontology generation, the BDI architecture enables adaptive, dynamic behavior, where it focuses its exploration on the most relevant aspects of the data while systematically searching for novel insights. BDI-based models have previously been applied in domains such as autonomous systems, robotics, and multi-agent coordination (Davoust et al., 2020; Gavigan & Esfandiari, 2022; Veres & Luo, 2004), demonstrating their ability to support intelligent, goal-driven behavior in complex environments. They have also been combined with LLMs in various domains, such as robotics (Frering, Steinbauer-Wagner, & Holzinger, 2025), emotion recognition (B. Xu et al., 2024), and linguistics (Silva et al., 2025).

The proposed BDI agent approach comprises an iterative process of document retrieval and reasoning, guiding an agent that autonomously generates questions, retrieves targeted information, and integrates new insights without requiring an exhaustive review of all available documents. In contrast to conventional retrieval systems, the agent's selection of documents is driven by its evolving beliefs and desires, allowing it to prioritize information that can contribute to covering the open question space. This approach moves beyond passive retrieval, enabling active and self-directed exploration that aims to balance the goals of novelty discovery, goal fulfillment, and semantic coherence (DeBellis, Dutta, Gino, & Balaji, 2024). Traditional ontology creation, on the other hand, relies on manual specification by domain experts. This process is often time-consuming, costly, error-prone, and impractical

when knowledge is constantly evolving or when domain experts are unavailable (Al-Aswadi, Chan, & Gan, 2020). This challenge is particularly pronounced in dynamic environments where knowledge is frequently updated, or where expertise is scarce or inaccessible.

This study contributes to formal reasoning research by introducing a dynamic, curiosity-driven agent for knowledge acquisition. Unlike methods such as Hearst pattern extraction, Rebel, or retrieval-augmented generation (Huguet Cabot & Navigli, 2021; DeBellis et al., 2024; Hearst, 1992), which passively extract facts based on prompts or input texts, our BDI-based model simulates an agent that actively and purposefully explores data, guided by goals and intrinsic curiosity. This enables autonomous refinement and expansion of knowledge structures. The approach also has practical value for businesses, offering a structured, reasoning-based method to extract insights from unstructured data. By aligning extracted knowledge with evolving goals and questions, the system supports adaptive, real-time decision-making, enabled by an explicit representation of the BDI model of an agent.

The chapter is structured as follows: Section 7.2 reviews related work; Section 7.3 introduces the BDI-based ontology generation model; Section 7.4 details the experimental setup; Section 7.5 reports the evaluation; Section 7.6 discusses the results; finally Section 7.7 concludes the chapter.

7.2 Related work

Ontology generation is the process of extracting and organizing knowledge from text sources. The goal is to identify and structure concepts, entities, and their relationships into a formal ontology (Gruber, 1993, 1995). This formal ontology allows for automated reasoning and knowledge representation (Gruber, 1995).

Earlier approaches to ontology generation include rule-based, pattern-based, and taxonomy-based methods. Rule-based methods have been suggested to extract conceptual relations from text corpora, often relying on linguistic rules and statistical techniques (Cimiano, Hotho, & Staab, 2005; Maedche & Staab, 2000; Völker, Hitzler, & Cimiano, 2007). Pattern-based methods identify recurring linguistic or syntactic structures in text to extract semantic relationships between concepts (Hearst, 1992). These methods often use predefined patterns or learn new ones (Agichtein & Gravano, 2000). Taxonomy-based approaches automatically extract and construct hier-

archical ‘is-a’ relationships between concepts to form structured knowledge representations (Navigli, Velardi, & Gangemi, 2003). These approaches use techniques like statistical methods, pattern extraction, and formal analysis to transform sources into formal taxonomies, forming the foundation for more complex ontologies (Ponzetto & Strube, 2007).

In the context of ontology generation, natural language processing (NLP) techniques are increasingly used to extract, organize, and structure knowledge from unstructured text. NLP methods such as named entity recognition (NER) (J. Li, Sun, Han, & Li, 2022), part-of-speech tagging (Chiche & Yitagesu, 2022), and dependency parsing (McDonald & Nivre, 2011) can be used to identify key concepts and relations within the text. These techniques allow for the automated creation of ontologies by extracting relevant entities and their semantic relationships.

The evolution of NLP has been influenced by the introduction of LLMs, which leverage pre-trained contextual knowledge to interpret linguistic context and meaning. As a result, LLMs have been explored for their potential in various stages of the ontology generation pipeline (Babaei Giglou et al., 2023; Joachimiak et al., 2024; Saeedizade & Blomqvist, 2024). Their capacity to reason over unstructured or semi-structured data, along with their ability to capture linguistic patterns, offer potential advantages in extracting and organizing domain-specific knowledge (Kommineni, König-Ries, & Samuel, 2024). Recent studies have investigated the use of LLMs in tasks such as constructing concept hierarchies (Joachimiak et al., 2024), generating knowledge graphs (Hofer, Obraczka, Saeedi, Köpcke, & Rahm, 2024), and supporting relation extraction through methods like retrieval-augmented generation (DeBellis et al., 2024) and the Rebel framework (Huguet Cabot & Navigli, 2021). These approaches indicate a growing interest in the potential of LLMs to support ontology development from text.

In this chapter, we build on the growing use of LLMs in ontology generation, as we propose an approach that replaces passive, corpus-wide extraction with iterative, goal-driven exploration guided by a cognitive agent model. We suggest the use of the BDI framework to support the generation of curiosity-based questions by LLMs. Unlike traditional and pattern-based methods that usually scan whole collections of texts, our approach focuses on finding new and relevant information based on specific goals (Hearst, 1992). This makes it possible to build ontologies more efficiently and ensures they better match the questions or topics of end users. Building on these developments, the following section introduces our BDI-guided architecture, detailing

how beliefs, desires, and intentions drive LLM-powered question generation, document retrieval, and ontology construction through an iterative reasoning loop.

7.3 BDI model

We propose a BDI-inspired model that enables reasoning-driven and iterative knowledge acquisition. The Belief-Desire-Intention (BDI) model suggests that intelligent behavior arises not only from immediate decisions, but from a structured interplay between the agent's beliefs about the world, its goals for the future, and the corresponding plan of action to achieve its objectives based on its current beliefs (Bratman, 1987; Kakas, Mancarella, Sadri, Stathis, & Toni, 2008). Beliefs contain the information the agent has about the world in which it is behaving. Desires are the actual states in the agent's environment that it wants to achieve. Then, intentions are the steps the agent will attempt to perform, guiding its planning and action execution (Rao & Georgeff, 1991). These different components allow the model to have autonomous decision-making with a long-term goal, while it can react immediately in a dynamic environment (Rao & Georgeff, 1995). To operationalize this framework for ontology generation, we design an agent that iteratively acquires knowledge by generating and executing information-seeking questions. These questions are based on its current beliefs, intentions, and desires, thereby reflecting information gaps or areas of low confidence. The questions are used to retrieve documents, after which triples are extracted to update the knowledge base. Through this cycle, the system iteratively builds, refines, and expands its knowledge base, without relying on predefined queries or exhaustive document review, as can be seen in Figure 7.1. Additionally, an example is provided in Table 7.1.

Table 7.1: Example of the BDI model

Step	**Output**
Initial query	What does the dataset reveal about companies?
Set desires	['Gain information on people in companies', 'Find information on the finances of companies']
	=== BDI Agent Starting ===
(1) Search query	Searching for: What does the dataset reveal about companies?

(2) Search returned results	Search returned the top *r* results
(3) Example result	The S.A. (corporation) Hypermarcas, which earns 1800000000 annually, has a subsidiary called Mantecorp.
(4) Extract RDF triples	(HYPERMARCAS, EARNS, 1800000000) (HYPERMARCAS, HAS_SUBSIDIARY, MANTECORP)
(5a) Update beliefs	HYPERMARCAS EARNS 1800000000
(5b) Update beliefs	HYPERMARCAS HAS_SUBSIDIARY MANTECORP
…	…
(6) Deliberate desires	Selected desire in deliberate: Find information on the finances of companies

(7) Form intention-question pairs	(Intention: Investigate the financial health of companies, New Query: What patterns can be observed in the financial data of various companies in the dataset?) (Intention: Identify common industries among the companies, New Query: Which industries are predominantly represented in the dataset?) (Intention: Understand the geographical distribution of these companies, New Query: In which countries or regions do most of the companies in the dataset operate?) (Intention: Examine the size and scale of these companies, New Query: What is the range of employee counts among the companies in the dataset?) (Intention: Investigate the corporate structure of these companies, New Query: Are there any common types of corporations or organizational structures among the companies in the dataset?)
(8) Select intention	Desire: Find information on the finances of companies Intention: Investigate the financial health of companies Query: What patterns can be observed in the financial data of various companies in the dataset? Priority: 0.76 Status: in progress
(9) Execute query	What patterns can be observed in the financial data of various companies in the dataset?
Search query	Step (1)

In the following sections, we describe the core components of the BDI agent: beliefs, desires, and intentions; document extraction; internal knowledge processing; intention-question generation; and intention selection and deliberation. We then discuss the ontology generation process. This is followed by the experimental setup, which outlines the models used, datasets, and evaluation procedures.

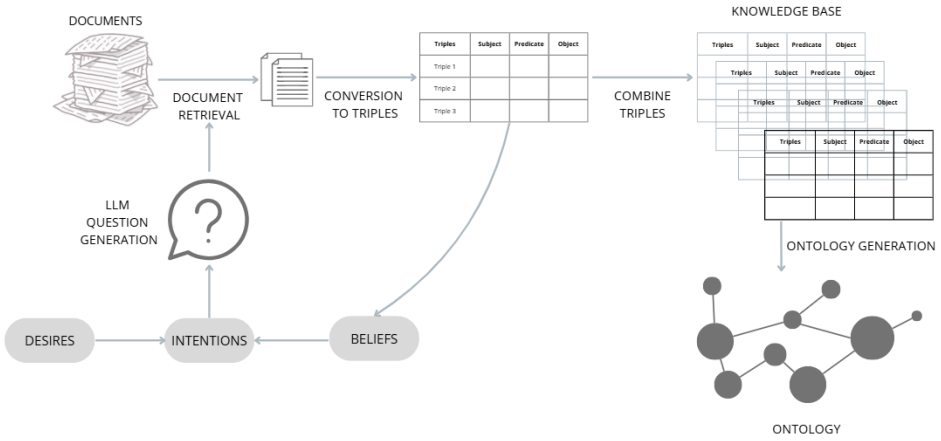

Figure 7.1: BDI agent for ontology generation.

7.3.1 Beliefs, desires and intentions

Beliefs ($\mathcal{B}$) represent the agent's current knowledge about the dataset, encoded as a set of RDF triples:

$$\mathcal{B} = \{(s, p, o) \mid s \in \mathcal{S}, p \in \mathcal{P}, o \in \mathcal{O}\}$$

where $\mathcal{S}, \mathcal{P}, \mathcal{O}$ are sets of subjects, predicates, and objects, respectively. All information is represented as complete triples; if a relation can not be grounded with a valid object, it is not included in the belief set. Beliefs are dynamically updated as new information is acquired.

Desires ($\mathcal{D}$) represent the agent's high-level goals that guide its behavior. In the context of knowledge extraction, a desire specifies a specific information need. We define the set of desires as:

$$\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$$

where each $d_i \in \mathcal{D}$ is represented as a structured prompt or textual formulation expressing an open-ended information-seeking goal. Examples of desires can be found in Section 7.4.3.

Intentions ($\mathcal{I}$) convert desires into concrete actions, bridging abstract goals and executable queries. Formally,

$$\mathcal{I} = \left\{ (d, q, \pi, \sigma, a, r) \;\middle|\; \begin{array}{r} d \in \mathcal{D},\ q \in \mathcal{Q},\ \pi \in \mathbb{R}_{\geq 0}, \\ \sigma \in \Sigma,\ a \in \mathbb{N},\ r \in \mathcal{R} \end{array} \right\}$$

i.e., $\mathcal{I}$ consists of six tuples of the form $(d, q, \pi, \sigma, a, r)$ where

- $d \in \mathcal{D}$: associated desire,
- $q \in \mathcal{Q}$: a generated question,
- $\pi \in \mathbb{R}$: priority score, based on novelty and semantic alignment,
- $\sigma \in \{\text{pending}, \text{in\_progress}, \text{completed}, \text{failed}\}$: execution status,
- $a \in \mathbb{N}$: number of attempts,
- r: results obtained from execution.

Priority π is computed as:

$$\pi = \text{novelty}(q, \mathcal{Q}_{prev}) \times \text{alignment}(d, q)$$

Here, novelty is computed using the cosine distance between question embeddings, and alignment is calculated as the average embedding similarity between the desired and the generated intention.

Intentions are discarded after exceeding a maximum attempt count a_{max} to avoid redundant querying.

7.3.2 Document extraction

The agent searches for relevant documents using vector embeddings and FAISS indexing (Douze et al., 2025). FAISS is a library for efficient similarity search over

dense vectors. This method retrieves documents based on semantic similarity to the question, filtering the results using a relevance threshold.

In more formal detail, building on the agent's beliefs $\mathcal{B}$, desires $\mathcal{D}$, and intentions $\mathcal{I}$ as defined above in Section 7.3, where:

- $\mathcal{Q}_c$ is the current question,
- $\mathcal{C} = \{doc_1, doc_2, \ldots, doc_n\}$ is the set of all documents,
- $\text{Embed}(x) \in \mathbb{R}^k$ is the embedding function that maps input x to a vector in k-dimensional space,
- the Euclidean distance between two vectors $a, b \in \mathbb{R}^k$ is defined as:

$$d(a, b) = \|a - b\|_2 = \sqrt{\sum_{j=1}^{k} (a_j - b_j)^2},$$

- $\theta \in \mathbb{R}_{\geq 0}$ is a relevance threshold on distance,
- $s_i \in \mathbb{N}$ is the number of times document doc_i has been retrieved in the past, and
- $\alpha \in \mathbb{R}_{\geq 0}$ is a penalization coefficient for repeated retrieval,

The adjusted similarity score for each candidate document doc_i is defined as:

$$\tilde{d}_i = d(\text{Embed}(\mathcal{Q}_c), \text{Embed}(doc_i)) \cdot (1 + \alpha \cdot s_i)$$

The set of relevant documents is:

$$\mathcal{C}_{\text{rel}}(\mathcal{Q}_c) = \left\{ doc_i \in \mathcal{C} \;\middle|\; \tilde{d}_i \leq \theta \right\}$$

FAISS is used to efficiently compute and retrieve the top k nearest neighbors by raw distance, with penalization of frequently retrieved documents to encourage novelty (Douze et al., 2025). Only documents whose adjusted distance $\tilde{d}_i$ remains below the threshold θ are retained as relevant.

7.3.3 Internal knowledge processing

From the set of relevant documents $\mathcal{C}_{\text{rel}}(\mathcal{Q}_c)$, the agent extracts structured information using an LLM-based extraction method. Each document is transformed into

a set of RDF-style triples that represent factual knowledge. Let $\mathcal{T}_{\text{extracted}}$ be the set of newly extracted triples.

The agent incorporates the newly extracted triples $\mathcal{T}_{\text{extracted}}$ into its existing belief base $\mathcal{B}$, resulting in an updated set of beliefs:

$$\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{T}_{\text{extracted}}$$

This union represents the continuous integration of new factual information to maintain an up-to-date knowledge state.

7.3.4 Intention-question generation

A central feature of the agent's reasoning process is the generation of intention-question pairs. Each intention I is formed by selecting relevant beliefs $\mathcal{B}$ and current desire d^*, where $d^* \in \mathcal{D}$ denotes the desire with the least amount of information in the current knowledge base, representing a knowledge gap that the agent seeks to close. The concepts of beliefs $\mathcal{B}$, desires $\mathcal{D}$, and intentions $\mathcal{I}$ are formally defined and elaborated in Section 7.3.1.

Formally, we define a function:

$$F : \mathcal{B} \times \mathcal{D} \to \mathcal{P}(\mathcal{I}),$$

where each intention $I \in \mathcal{I}$ includes, among other fields, an associated question $q \in \mathcal{Q}$. Thus, the mapping from beliefs and desires to questions can be expressed by composing with a projection:

$$G(I) = I_q, \quad \text{and} \quad G \circ F : \mathcal{B} \times \mathcal{D} \to \mathcal{P}(\mathcal{Q}).$$

In implementation, this two-step process is approximated by a single LLM prompt, which directly produces multiple intention-question pairs conditioned on the current beliefs, the selected desire, and recent query history. Each intention includes its corresponding question, eliminating the need to apply a separate function G after generation. The LLM prompt thus encapsulates the logic of both F and G, co-generating intentions and their corresponding ontology-building questions.

The questions q serve as the primary interface between the agent and its environment, enabling active information acquisition and belief updates. These questions

are dynamically generated via a prompt-based mechanism grounded in the agent's current knowledge $\mathcal{B}$ and goal d^*.

The prompt guiding the generation of intention-question pairs includes the following constraints:

- Do not hallucinate: only provide themes based on the actual content of the sentences.
- No theoretical data: avoid introducing fictitious examples or extrapolated information.
- Session-limited learning: do not use data from this session for learning beyond the task or retain any data outside of this session.
- Formulate intentions based on the current desire d^*.
- Formulate exactly ONE question PER intention, but you can formulate multiple intention-question couples.
- Intentions can be novel and creative in relation to the desire d^*.
- Formulate broad, ontology-building questions, avoiding overly specific examples.
- Avoid questions semantically close to previous questions $\mathcal{Q}_{\text{prev}}$.

The output structure is thus a set of intention-question pairs:

$$\{(I_i, q_i)\}_{i=1}^{m}$$

where each q_i is generated conditioned on the current beliefs $\mathcal{B}$, current desire d^*, and previous questions $\mathcal{Q}_{\text{prev}}$.

7.3.5 Intention selection and deliberation

At each cycle, the agent first attempts to select the next intention I^* to execute. Eligible intentions are those pending execution and which have not exceeded the maximum attempt limit A_{max}:

$$\mathcal{I}_{\text{eligible}} = \{I_j \in \mathcal{I} \mid s(I_j) = \text{pending} \wedge a(I_j) < A_{\text{max}}\}$$

The agent selects the highest priority intention:

$$I^* = \arg\max_{I_j \in \mathcal{I}_{\text{eligible}}} p(I_j)$$

If no eligible intentions exist, the agent enters the deliberation phase to select a new desire d^* to focus on, prioritizing desires with the least associated information:

$$d^* = \arg\min_{d \in \mathcal{D}} |\mathcal{B}_d|.$$

The agent continues deliberating and generating intentions for d^* as long as:

- There exists at least one pending intention $i \in \mathcal{I}_{\text{pending}}$ with priority $\geq \delta$, or
- Not all desires in $\mathcal{D}$ have been fully explored.

An iteration limit $\kappa \in \mathbb{N}$ ensures the process terminates after a fixed number of cycles.

7.3.6 Ontology generation

After the BDI reasoning cycle terminates, all triples collected during execution are consolidated into a structured ontology. The transformation is guided by a structured prompt that enforces naming conventions and semantic best practices, and returns the ontology in `Turtle (.ttl)` format. The prompt specifies:

- Classes start with uppercase (e.g., `Company`).
- Properties start with lowercase (e.g., `hasIndustry`).
- Use of semantic relationships with a defined domain and range.
- `@prefix` notation for readability.

The next step is the experimental setup used to assess our BDI agent's ability to generate ontologies across diverse domains, using a standard benchmark dataset and state-of-the-art evaluation metrics.

7.4 Experimental setup

7.4.1 Large language models

In this study, we utilize the Mistral-7B-Instruct-v0.3 model and OpenAI's GPT-3.5 model. Mistral is used for triple extraction and generation of intention-question pairs, selected for its efficiency on GPU-supported laptops (Jiang et al., 2023). For ontology construction from extracted triples, we use the GPT-3.5 model. Different combinations of tools and LLMs are possible, but in this context, we have chosen these LLMs as tools to demonstrate the feasibility of our BDI-based approach. A systematic comparison of alternative LLMs using this framework is beyond our present scope and is left for future work.

7.4.2 Data

For this study, the Text2KGBench dataset was employed (Mihindukulasooriya, Tiwari, Enguix, & Lata, 2023). The Text2KGBench is a benchmark dataset created to evaluate how well models convert natural language into structured knowledge graphs. The dataset consists of texts paired with annotated knowledge graphs, where each graph captures entities, relations, and types extracted from the unstructured input documents. Text2KGBench is composed of two subsets: Wikidata-TekGen and DBpedia-WebNLG. In this study, we focus on the DBpedia-WebNLG dataset, which reuses the alignments created in the webNLG corpus. The DBpedia-WebNLG subset consists of 19 categories, from which we select the categories airport, company, and written work. These categories were chosen for their combination of ontological clarity and diversity of entities and relationships, enabling a robust assessment across different graph construction challenges. The number of classes, relations, and documents per category can be found in Table 7.2.

Table 7.2: Overview of ontology categories.

Ontology	Classes	Relations	Documents
Airport	14	39	306
Company	10	28	153
Written work	10	44	322

7.4.3 Implementation details

The BDI agent is initialized with user-provided desires and the initial question for each ontology category. For the airport category, the desires are: 'Gain information on the facilities of airports', 'Gain information on the locations of the airports'. The initial question is: 'What does the dataset reveal about the airports?'. For the company category, the desires are: 'Gain information on the people in companies', 'Gain information on the finances of companies'. The initial question is: 'What does the dataset reveal about the companies?'. For the written work category, the desires are: 'Gain information on the different authors', 'Gain information on the texts'. The initial question is: 'What does the dataset reveal about the written works?'.

The system operates under a set of fixed parameters to ensure consistent behavior across all experimental conditions. These parameters can be found in the corresponding code. The novelty threshold θ was set to 1.8, determining the minimum required dissimilarity for a newly generated question to be considered sufficiently distinct from previous queries. The priority scaling factor α was assigned a value of 0.1 for the Company domain and 0.05 for both Airport and Written work, as these settings yielded the best empirical performance during development. Document retrieval was performed in two stages: initially selecting the top $k = 30$ documents based on relevance, from which the top $r = 10$ were subsequently used for belief integration. A maximum of $\kappa = 30$ deliberation cycles was permitted per run, and the maximum number of execution attempts per intention was limited to 3. To regulate belief acquisition, the system allowed at most 10 belief updates before re-entering the deliberation phase.

7.4.4 Baseline methods

For comparison, we evaluate our BDI agent against two established information extraction approaches: Rebel and Hearst pattern-based extraction. The Hearst method (Hearst, 1992) leverages predefined lexico-syntactic patterns to identify semantic relations within text corpora. This pattern-based approach serves as a classical baseline for structured knowledge extraction from raw text. Rebel (Huguet Cabot & Navigli, 2021) is a retrieval-augmented system that extracts relational facts from unstructured text using large language models. It operates by passively processing input documents to identify entities and relations without explicit goal-driven behavior.

7.4.5 Evaluation

For this study, graph and continuous precision, recall, and F1-scores are used, as proposed by Lo et al. (2024). Graph precision, recall, and F1-scores are evaluation methods that compare two graphs based on the semantic and structural similarity of their nodes (Lo, Jiang, Li, & Jamnik, 2024). Each node is first embedded using a pre-trained language model, and these embeddings are then passed through a simple graph convolution with $K=2$ to incorporate information from the node's local neighborhood. Pairwise cosine similarities between the resulting node embeddings from the predicted and reference graphs are computed. The best one-to-one alignment between nodes is determined using the Hungarian algorithm, and graph precision and recall are calculated as the average similarity over matched nodes, normalized by the number of nodes in the predicted and reference graphs, respectively. The Graph F1-score is then computed as the harmonic mean of precision and recall.

Continuous precision, recall, and F1-scores evaluate edges based on the semantic similarity of the nodes that make up each edge. The minimum cosine similarity between the source and target node embeddings determines edge similarity. The best edge alignment is computed using the Hungarian algorithm, and precision and recall are derived based on the aligned edges. The F1-score is then calculated as the harmonic mean of continuous precision and recall.

7.5 Results

This section evaluates the effectiveness of the BDI model in constructing ontologies from unstructured data. For brevity, we report results for the best-performing configuration of the BDI model only. Table 7.3 shows a comparative overview of performance across the three categories. To illustrate the model's behavior, we include the visualizations of the graphs of the company category in Figure 7.2, and the visualizations of the graphs of the airport and written work categories, together with the examples of questions generated during the ontology construction process in Appendix F till J.

Table 7.3: Evaluation metrics for airport, company, and written work categories.

Ontology category	Method	Cont. precision	Cont. recall	Cont. F1	Graph precision	Graph recall	Graph F1	Number of documents	Number of questions
Airport	Hearst	0.034	0.436	0.064	0.068	0.551	0.121	306	-
	Rebel	0.167	0.363	0.229	0.255	0.415	0.316	306	-
	BDI	**0.183**	**0.516**	**0.270**	**0.265**	**0.686**	**0.382**	88	16
Company	Hearst	0.294	**0.537**	0.380	0.378	**0.566**	0.453	153	-
	Rebel	0.301	0.409	0.347	0.430	0.488	0.457	153	-
	BDI	**0.356**	0.497	**0.415**	**0.452**	0.558	**0.499**	98	30
Written work	Hearst	0.068	0.437	0.118	0.136	**0.564**	0.220	322	-
	Rebel	0.211	0.407	0.278	0.344	0.496	0.407	322	-
	BDI	**0.294**	**0.443**	**0.353**	**0.399**	0.504	**0.445**	116	30

Note. Best values per category in each column are shown in **bold**.

7.5.1 Airport

For the Airport category, the BDI model outperforms both the Rebel and Hearst models across most evaluation metrics. On the content level, the BDI model achieves a precision of 0.183, a recall of 0.516, and an F1-score of 0.270, compared to Rebel's precision of 0.167, a recall of 0.363, and an F1-score of 0.229, and Hearst's precision of 0.034, a recall of 0.436, and an F1-score of 0.064. Similarly, at the graph level, the BDI model attains a precision of 0.265, a recall of 0.686, and an F1-score of 0.382, outperforming Rebel's 0.255, 0.415, and 0.316, as well as Hearst's 0.068, 0.551, and 0.121. Notably, the BDI agent achieves this performance using only 88 documents.

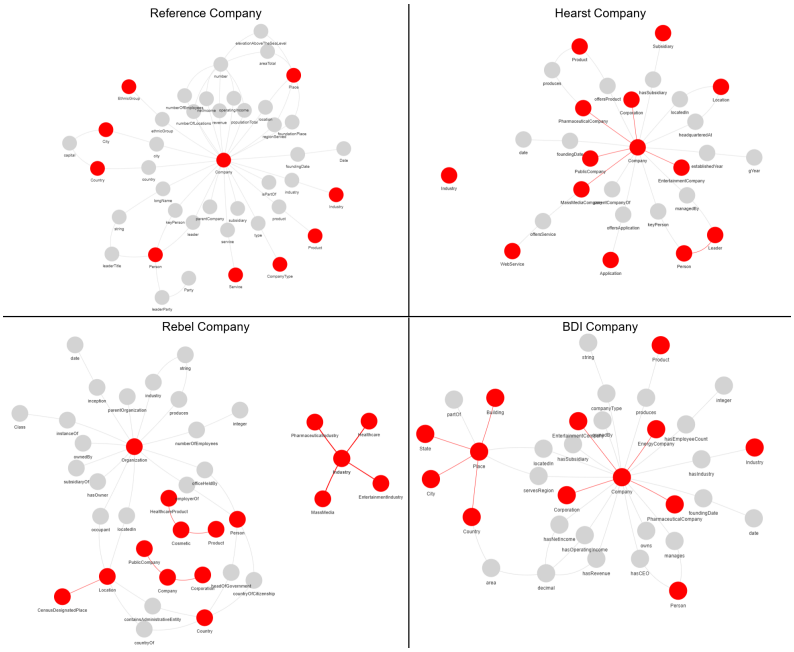

Figure 7.2: Graph visualization of the company category for the reference, Hearst, Rebel, and BDI ontology.

Note. The red dots indicate the ontology's concepts (classes), while the grey nodes are supporting elements, such as properties, instances, or domains.

7.5.2 Company

For the Company ontology, the Rebel method achieves a continuous precision of 0.301, a recall of 0.409, and an F1-score of 0.347. Its graph-level precision, recall, and F1-scores are 0.430, 0.488, and 0.457, respectively. The Hearst method performs similarly, with a continuous precision of 0.294, recall of 0.537, and F1-score of 0.380, and graph-level precision, recall, and F1-scores of 0.378, 0.566, and 0.453, respectively. In comparison, the BDI agent improves upon these results with a continuous precision of 0.356, recall of 0.497, and F1-score of 0.415. At the graph level, the BDI agent attains a precision of 0.452, a recall of 0.558, and an F1-score of 0.499, demonstrating consistent gains across all metrics. The BDI agent achieves this performance using only 98 documents and 30 questions, in contrast to Rebel and Hearst, which both use 153 documents.

7.5.3 Written work

For the Written Work ontology, the Rebel method achieves a continuous precision of 0.211, a recall of 0.407, and an F1-score of 0.278. Its graph-level precision, recall, and F1-scores are 0.344, 0.496, and 0.407, respectively. The Hearst method performs lower in continuous precision at 0.068 but achieves recall and F1-scores of 0.437 and 0.118, with graph-level precision, recall, and F1-scores of 0.136, 0.564, and 0.220, respectively. The BDI agent outperforms both Rebel and Hearst with a continuous precision of 0.294, recall of 0.443, and F1-score of 0.353. At the graph level, the BDI agent attains a precision of 0.399, a recall of 0.504, and an F1-score of 0.445, showing improvements across all evaluated metrics. The BDI agent achieves this performance using only 116 documents and 30 questions, in contrast to Rebel and Hearst, which both use all 322 documents.

7.6 Discussion

The BDI agent consistently outperforms the Rebel and Hearst models on all data categories. Thereby, we establish a foundation for dynamic, goal-driven knowledge acquisition using the BDI framework.

Building on the benchmarked results presented in this chapter, the BDI agent approach could next be applied to organizational service data, such as reviews, training

manuals, or other sets of organizational documents (Osman et al., 2022). Within the broader scope of this dissertation, this would allow AI to structure unorganized information and link customer-facing issues to internal knowledge resources, creating a feedback loop between frontline experiences and organizational knowledge. Ontology generation thus serves as a macro-level complement to the micro-level analyses of customer interactions explored in earlier chapters.

For instance, in a customer service call center, clients frequently contact the firm with various inquiries. A BDI-driven ontology agent can explore extensive collections of internal resources, such as troubleshooting documents, past tickets, and policy updates (Alaimo & Kallinikos, 2021). Rather than requiring agents to search manually, the system dynamically maps a customer's complaint to the most relevant knowledge structures, surfacing both technical solutions and organizational guidelines. By making internal knowledge directly accessible at the point of service, the approach closes the loop between frontline interactions and firm-level resources, illustrating how ontology generation can support more effective service delivery and organizational learning.

Future work may explore other datasets, especially larger datasets, to test the performance of the BDI model. Investigating how targeted question generation could allow the model to focus on specific parts of these larger datasets may improve efficiency while maintaining high recall. Similarly, future work could explore the impact of using larger or domain-specific models to assess how model capacity and specialization affect the quality of extracted knowledge and generated ontologies.

Future research could look into automatic initial question and automatic desire generation, which introduces additional challenges such as goal selection, redundancy detection, and semantic drift. It could also explore coupling question generation with ontology construction, allowing the ontology to be built incrementally as questions are posed and answered. This approach would enable a tighter integration between reasoning and knowledge representation. Additionally, introducing constraints and pruning mechanisms in ontology construction may enhance scalability and relevance.

This work focuses primarily on performance and effectiveness, and does not include a detailed evaluation of computational efficiency or energy consumption. While the system incorporates adaptive querying, a systematic analysis of energy usage and resource requirements across methods remains outside the scope of this study. We consider this an essential direction for future work, particularly for applications

requiring deployment at scale.

Finally, research could explore ways to adjust the focus and depth of questions, enabling the agent to transition from broad topics to specific details based on what is most needed. Research could also explore multi-agent systems, where different agents specialize in various subdomains or perspectives and collaborate to create more comprehensive ontologies. Finally, future work could examine how the ontology evolves as more data is added or when additional questions or needs arise, automatically adjusting to stay organized and effective without human oversight.

7.7 Conclusion

We present a BDI-based system for ontology generation that integrates selective inquiry and curiosity-driven reasoning into a dynamic, goal-oriented pipeline. Unlike traditional methods, the BDI model prioritizes relevant data through targeted questions, simulating human-like reasoning. The BDI model structurally outperforms the Rebel and Hearst baseline approaches. The results highlight the BDI model's ability to build comprehensive ontologies without exhaustive data processing, offering an efficient alternative for large-scale tasks. By focusing on dynamic, curiosity-driven learning, the BDI model adapts to evolving knowledge needs, providing a flexible solution for complex domains.

Chapter 8

General discussion and conclusion

8.1 Introduction

The transformation of the global economy toward service-dominance has changed how value is created through customer-firm interactions (Grönroos, 2011). Service encounters present critical moments of co-creation where customers and service providers jointly shape service outcomes (Vargo & Lusch, 2008). Millions of these interactions occur daily across various channels (World Bank, 2024), containing rich, multi-layered information including affective states, strategic communication patterns, performance indicators, and structured knowledge about service processes. The extraction of this information can help service firms enhance the service delivery process and improve outcomes (M. Blaurock et al., 2024; Libai et al., 2022).

However, a central theoretical problem remains underexplored: current service management and information-processing theories provide limited guidance on how organizations can systematically capture and utilize the complex, multi-dimensional information embedded in high-volume, human-to-human service interactions. While service-dominant logic emphasizes value co-creation through information exchange (Vargo & Lusch, 2008), existing frameworks do not specify how the multiple layers of service interaction, from explicit content to implicit emotional cues and aggregated patterns, can be operationalized as organizational knowledge at scale (Choi, 2018; Kumar et al., 2013). The scale and complexity of human-to-human service interactions make manual analysis impractical, leaving much valuable knowledge hidden in conversational data (H. Chen et al., 2012; McKinsey & Company, 2024). This gap represents a critical opportunity to advance service theory through frameworks that systematically extract and utilize multi-layered conversational knowledge as a core organizational capability.

This dissertation addresses this theoretical challenge by investigating how AI approaches can systematically convert tacit conversational knowledge into explicit organizational assets, reconceptualizing service interactions as multi-layered information structures that can be analyzed and implemented across content, implicit, and aggregated layers. Through six interconnected studies, this research developed and compared various AI approaches to systematically identify patterns across different layers of service interactions and generate actionable insights from conversational data. This work positions AI-driven extraction capabilities as fundamental organizational resources, not just applied tools. It contributes to service management theory and information systems theory while addressing practical human-AI collaboration

challenges in service delivery (Bardhan et al., 2010).

The conceptual framework introduced in Chapter 1 positioned this research at the intersection of customer service interactions, multi-layered information structures, and artificial intelligence technologies. This framework proved instrumental in structuring the investigation across the three identified layers: explicit content, implicit signals, and aggregated patterns, while guiding the systematic comparison of AI approaches. The convergence of these three domains provided the theoretical foundation for understanding how conversational knowledge can be transformed into organizational capabilities through AI-driven extraction processes.

The findings suggest three main theoretical advances: how organizations can systematically extract multi-layered information from service interactions, what capabilities are required to decode conversational structures, and how AI-driven processing enhances organizational learning from these encounters. These contributions span theoretical advancement in service management, computer science, and practical applications for service organizations.

This chapter is organized as follows. First, a synthesis of the key results across the six studies is presented. Subsequently, this chapter addresses the research questions that guided this dissertation. The discussion then explores the theoretical implications of the intersection of service management and AI, followed by practical contributions for service organizations. The chapter then addresses ethical and societal implications of AI-driven information extraction systems, discusses research limitations and future work avenues, and concludes with a summary of the dissertation.

8.2 Summary of the chapters

In Chapter 2, a systematic literature review of 99 empirical studies revealed six central themes of technology integration in human-to-human service interactions: pre-service optimization, interaction intelligence, service agent well-being, service agent monitoring, emotion work, and augmentation. The thematic mapping, anchored in socio-technical systems theory, demonstrated how core technological capabilities, including AI, can effectively augment rather than replace service agents across various stages and tasks in customer service interactions. This comprehensive framework contributes to the theoretical foundation of this dissertation, demon-

strating how technology-driven capabilities can augment human expertise throughout service delivery. It offers insights for socio-technical systems theory and helps to develop a conceptual foundation for understanding human-AI collaboration in service contexts. It thereby contributes to an understanding of how technology can augment service agent effectiveness, efficiency, and well-being, leading to improvements in the service process.

Chapter 3 presented a comparative evaluation of neural network architectures for automatic speech emotion recognition in real-life customer service conversations. The study demonstrated that multi-layer perceptrons, one-dimensional convolutional neural networks, and neural machine translation models all outperformed baseline classifiers. This performance difference suggests there are meaningful patterns in the data relating to emotion labels. While the neural machine translation model with an attention mechanism achieved the highest F1-score, statistical analysis revealed no significant performance differences among the three neural network architectures. Therefore, the results show the multi-layer perceptron as the optimal choice due to its simplicity while maintaining competitive performance. These findings contribute to multi-layered information processing as they demonstrate how AI can systematically decode emotional structures embedded in the implicit layer of service communication.

Chapter 4 explored machine learning techniques for predicting customer dissatisfaction from voice-to-voice service interactions through multimodal integration of textual and auditory signals. The chapter demonstrated that cross-attention mechanisms significantly outperformed audio-only, text-only, and late-fusion approaches for dissatisfaction detection. Grounded in communication theory's interactivity principle, the findings revealed that optimal integration of verbal and vocal components requires more complex attention mechanisms that can capture complex interdependencies between modalities. These findings advance AI-driven extraction capabilities by demonstrating how multimodal approaches can decode dissatisfaction signals from the implicit layer of service conversations, providing service organizations with enhanced knowledge on customer dissatisfaction for proactive intervention strategies and service recovery.

In Chapter 5, AI models were employed to extract response strategies from firm social media interactions with consumers. Seven distinct response strategies were extracted through a multi-label classification approach, recognizing that multiple response strategies can be deployed within a single conversation. The comparative analysis between deep learning and large language models revealed that while LLMs

with examples performed reasonably well, a custom-trained multi-layer perceptron model using bag-of-words representations achieved the best performance across different embedding types. These findings demonstrate how AI can identify communication patterns in service interactions, providing organizations with a systematic understanding of their strategic communication approaches. It also contributes to a foundation for developing AI tools that can guide service agents in selecting appropriate response strategies, contributing to more effective and consistent customer engagement across digital service channels. It demonstrates how AI-driven extraction capabilities can systematically analyze conversational structures to identify strategic patterns from the content layer of service encounters, which reveals firms' information co-creation processes and advances service-dominant logic.

Chapter 6 discussed the potential of LLMs to support human evaluators in identifying potential errors in call quality monitoring forms. The three-step comparative approach involved comparing model assessments with primary human evaluator decisions, validating discrepancies through blind secondary human reviews, and analyzing patterns in cases where LLMs flagged potential human errors. The findings revealed both advantages and limitations of integrating LLMs into quality monitoring processes, providing insights into when models successfully identify human oversight versus when they underperform. The results demonstrate the value of human-AI collaboration in quality monitoring, showing how AI can enhance consistency and reliability while preserving human agency in complex evaluation tasks. These insights contribute to understanding the effectiveness of LLMs as supportive tools in call quality monitoring. The findings highlight scenarios where AI adds value while recognizing the importance of human judgment in complex evaluation tasks, advancing theoretical understanding of when AI complements versus when it falls short of human expertise.

Finally, Chapter 7 extends the dissertation's multi-layered view beyond the conversational domain. Using a belief-desire-intention (BDI) agent, the system dynamically generates questions, retrieves relevant documents, and incrementally organizes insights into an ontology, creating a goal-directed, curiosity-driven approach to knowledge extraction. Evaluation demonstrated that the BDI model outperformed baseline methods across continuous and graph-based evaluation metrics. This ontology-based approach shows how AI can also structure the organizational knowledge resources that underpin those encounters. In doing so, it bridges micro-level interaction analysis with macro-level knowledge management, indicating how customer com-

plaints, dissatisfaction signals, or response strategies can be systematically linked back to internal resources, such as troubleshooting, policies, and training materials. This chapter thereby demonstrates AI's support for effective service delivery and organizational learning via an integration of insights from aggregated organizational knowledge resources.

8.3 Addressing the research questions

This dissertation investigated the application of AI in customer service interactions through six interconnected studies. It contributes to bridging a critical gap between the rich information potential embedded in service encounters and organizations' ability to extract and utilize this multi-layered knowledge systematically. The findings from these studies collectively address the research questions established in Chapter 1 and demonstrate how the conceptual framework (Figure 1.1) guides the systematic extraction of meaningful insights from customer service interactions. Across different modalities and analytical layers, content, implicit signals, and aggregated patterns, these results illustrate how AI can operationalize the framework, while also contributing to the theoretical understanding of AI applications in service management.

The main research question asked: "*How can artificial intelligence be employed to extract conversational patterns and information from customer service interactions?*". Through the progression of the six studies presented in this dissertation, a comprehensive framework emerges that demonstrates AI's capacity to operate across multiple dimensions of customer service analysis. The research reveals that AI employment requires understanding service encounters as multi-layered information sources, encompassing content, implicit, and aggregated levels. This multi-layered approach spans different interaction modalities, including speech, text, and social media platforms, while addressing various analytical levels from emotional recognition to strategic understanding and quality monitoring. This approach contributes to service management theory, as it provides evidence that service encounters contain systematic information structures that can be computationally analyzed. It informs understanding of how value co-creation processes can be enhanced through systematic information extraction that integrates the perspectives of customers with those of the service agents and firms (Grönroos, 2011). Thereby, these insights offer new perspectives on information processing within service-dominant logic (Vargo

& Lusch, 2008). The findings also suggest that AI can function both as a practical tool and as an analytical approach for uncovering patterns in service communication, thereby enhancing organizational information processing capabilities.

The first research question was: "*What artificial intelligence approaches can effectively extract information from customer service interactions?*". From a methodological perspective, this question addresses the challenge of selecting appropriate AI methodologies to analyze the complex communication structures present in service encounters. This question also offers insights into the field of AI-driven service management by examining how different computational approaches may reveal characteristics of service data and communication patterns. Throughout the period these studies were executed, the AI field has evolved rapidly, with successive innovations leading to increasingly large and complex models. As a result, this work examined a range of models of varying complexity, reflecting the field's progression. These include baseline classifiers (Chapter 3), multi-layer perceptron models (Chapters 3 and 5), more complex neural network models (Chapters 3, 4, and 5), neural machine translation models (Chapter 3), and large language models (Chapters 5, 6, and 7).

The research, as demonstrated throughout chapters 2 to 7, shows that AI models can be used to extract information from customer service interactions. These include emotional states, strategic response patterns in social media exchanges, discrepancies in quality assessments between human evaluations, and dynamic knowledge structures generated through ontology construction. Notably, the findings indicate that the most complex and computationally expensive models do not necessarily outperform simpler, smaller models (Chapters 3, 5). This finding suggests that simpler models can sometimes be sufficient to capture the relevant patterns in the available data. With the growing availability of increasingly complex models, selecting the most suitable one for a given task becomes crucial, as choices involve trade-offs in cost, energy consumption, and interpretability, particularly in applied contexts such as customer service analysis. In a similar regard, custom-trained models outperformed general-purpose pretrained ones (Chapters 3, 5), underscoring the value of domain-specific adaptation and suggesting that service interactions contain unique information structures that benefit from specialized analytical approaches. These findings establish how organizations can develop computational information processing capabilities that transform tacit conversational knowledge into explicit organizational assets, contributing to AI-driven service management.

The second research question focused on: "*How can artificial intelligence be used*

to analyze conversational structures and identify meaningful patterns in customer service interactions?". Understanding conversational structures requires AI systems to navigate the inherent complexity of human communication, which operates across multiple dimensions simultaneously. Customer service interactions are not merely exchanges of information. They are complex social phenomena involving various layers of emotional states, strategic intentions, and contextual nuances that unfold over time. This complexity is demonstrated through the foundational work that established how AI can augment human service agents by understanding the current technological landscape (Chapter 2).

The studies in this dissertation demonstrate that successful conversational analysis requires AI to be implemented across analytical approaches and data types. Customer service communication encompasses multiple information channels, from spoken dialogue with tone and emotion to written text with linguistic structures, and social media exchanges with unique public characteristics. Speech emotion recognition exemplifies this multimodal challenge, revealing AI's ability to decode emotional patterns embedded in vocal signals (Chapter 3), the integration of audio and text analysis that showed how combining modalities enhances understanding of customer dissatisfaction patterns (Chapter 4), and through strategic response pattern identification that demonstrated AI's capacity to recognize deliberate communication strategies in text-based social media interactions (Chapter 5).

Each signal presents unique challenges for pattern identification, necessitating the adaptation of AI systems accordingly. This adaptability is further demonstrated through quality monitoring applications that revealed how AI can analyze conversational patterns in human evaluation processes (Chapter 6), and through dynamic ontology generation that illustrated AI's potential for building comprehensive knowledge structures that capture the complexity of conversational relationships and patterns at higher levels over documents (Chapter 7). These applications collectively demonstrate that AI's conversational analysis capabilities extend from immediate pattern recognition to complex relationship modeling and knowledge construction. It shows how organizations can develop computational capabilities to systematically process multi-dimensional service information structures and enhance organizational learning from conversational interactions (Bardhan et al., 2010).

The third research question posed: "*What insights can artificial intelligence-driven information extraction provide to service firms for understanding service interactions?*". It examines how AI can generate actionable organizational knowledge and enhance

understanding of service processes. This information is systematically extracted, which shows that AI can transform tacit knowledge from service encounters into explicit organizational assets. Beyond data extraction, AI reshapes how organizations interpret content, implicit cues, and aggregated patterns, enabling proactive understanding, improved service agent performance and well-being, and more effective service strategies. These insights reveal patterns that human analysis alone might miss, while underscoring the continued importance of human judgment and empathy in service delivery.

The dissertation demonstrates that AI-extracted insights can be tailored to various stakeholder perspectives, offering unique value at multiple levels. First, the conceptual understanding that AI should augment rather than replace human service agents (Chapter 2) establishes a collaborative framework that enhances human capabilities while preserving the essential human elements of service delivery and contributes to the understanding of socio-technical systems in service contexts. Second, from the customer's perspective, AI provides insights into emotional states and dissatisfaction patterns. Speech emotion recognition reveals the affective dimensions of the customer experience (Chapter 3), and multimodal dissatisfaction analysis demonstrates how combining different data sources creates a more comprehensive understanding of customer states (Chapter 4).

Finally, from both service agent and organizational perspectives, AI insights reveal patterns that inform decision-making and drive performance improvement. Strategic response pattern identification in social media interactions enables organizations to gain a systematic understanding of their communication approaches and their effectiveness (Chapter 5). Quality monitoring applications reveal inconsistencies and potential improvements in evaluation processes, supporting more reliable and fair assessment systems (Chapter 6). At the highest level, dynamic ontology generation demonstrates how AI can build comprehensive knowledge structures that capture the full complexity of service interactions (Chapter 7). These organizational-level insights suggest that AI-driven extraction helps transform tacit service knowledge into more accessible organizational learning capabilities.

AI-based information extraction adds both theoretical and practical value by systematically converting tacit knowledge from service encounters into explicit organizational assets, supporting service-dominant logic's emphasis on value co-creation (Grönroos, 2011; Vargo & Lusch, 2008). Rather than treating interactions as isolated events, AI links them to reveal patterns, relationships, and trends visible only at scale

(Keith et al., 2004). This broader perspective allows firms to understand how individual interactions shape customer journeys, inform service agent development, and drive organizational learning over time. Consequently, analysis shifts from a reactive, event-focused approach to an integrated, forward-looking view, enabling more effective service design and delivery.

8.4 Theoretical contributions

The results of the six studies in this dissertation lead to various theoretical contributions across multiple domains, advancing both service management theory and artificial intelligence research. Broadly, the research shows that AI can systematically extract and operationalize multi-layered information from service interactions. It demonstrates how human-AI collaboration can enhance service processes without replacing human capabilities and how multimodal analysis deepens understanding of customer emotions and dissatisfaction. Finally, the dissertation illustrates how interdisciplinary approaches bridge service management and AI to uncover structured patterns in conversational data that neither field could fully identify independently. These contributions are discussed in detail in the subsections below.

8.4.1 Contributions to service management

First, this dissertation advances service management theory by demonstrating how AI-driven extraction capabilities can systematically process the multi-layered information embedded in service interactions. While prior research, particularly service-dominant logic, has emphasized value co-creation as a process involving multiple actors, it has focused mainly on who participates rather than the information content of interactions. This dissertation addresses this gap by reconceptualizing service encounters as multi-layered information structures, encompassing explicit content, implicit signals, and aggregated interaction patterns, thereby providing a systematic framework for operationalizing informational outcomes of service interactions. This provides systematic methods for analyzing information exchange processes that facilitate value co-creation, extending service-dominant logic (Grönroos, 2011), and contributing to the theoretical gap regarding how firms can operationalize the informational outcomes of service interactions.

Second, this dissertation contributes to the service management field, as it provides a comprehensive overview of technology's role in customer service through augmentation, as presented in Chapter 2. Grounded in socio-technical systems theory, the research demonstrates how combining AI with human service agents enhances human-centered services (Davies et al., 2017). This approach moves beyond viewing technology as a replacement for human capabilities and instead positions it as a collaborative partner that creates shared value (Grönroos, 2011). The findings support a framework for collaborative intelligence in service contexts, where human intuition and AI pattern recognition work together to enhance service delivery (J. J. Kim et al., 2025).

Third, this dissertation advances the understanding of affective information processing in service management. It demonstrates how AI can systematically extract emotional and dissatisfaction signals from multimodal service interactions. The research progresses from single-modal emotion recognition using audio signals to sophisticated multimodal integration of audio and textual data for dissatisfaction prediction. This contributes to service management theory by showing how information from the implicit layer of service conversations can be analyzed to enhance understanding of customer emotional states and satisfaction levels. The findings demonstrate that multimodal approaches significantly outperform single-modal methods, advancing theoretical understanding of how different communication channels carry complementary affective information in service contexts. Systematic methods for combining verbal and vocal signals to decode complex emotional patterns extend affective computing insights in service management, thereby enabling organizations to improve understanding and responding to customer affective states (Grandey et al., 2004).

Fourth, this dissertation demonstrates how AI can work alongside human evaluators to enhance quality monitoring processes, contributing to service quality theory. Rather than proposing AI as a replacement for human supervisors, the research shows how computational approaches can supplement human judgment while preserving human agency in quality assessment. This advances service quality theory (Folan & Browne, 2005) as it demonstrates how human-AI collaboration can improve consistency and accuracy in quality monitoring systems. The theoretical framework positions AI as an aid to both service agents and supervisors, creating more reliable and consistent quality monitoring systems while preserving human agency in decision-making.

Finally, this dissertation advances understanding of organizational learning in ser-

vice contexts by demonstrating how AI-driven extraction enables large-scale analysis of service interactions, uncovering phenomena previously difficult to observe (Marinova et al., 2016). AI extends traditional knowledge management approaches and enhances organizational information processing capabilities in service contexts while systematically transforming tacit knowledge embedded in service conversations into explicit organizational knowledge (Choi, 2018). This collaborative approach highlights the potential for AI to generate actionable insights from conversational data, contributing both to theoretical understanding and practical applications in service management.

8.4.2 Contributions to artificial intelligence literature

The dissertation also makes contributions to the fields of AI and machine learning through systematic model comparison and methodological innovation in real-world service contexts. This comparative approach is theoretically important because it challenges assumptions about model performance and provides empirical evidence for context-appropriate model selection (Ding, Tarokh, & Yang, 2018). The research demonstrates how systematic evaluation of AI models across service tasks can reveal fundamental properties about the computational requirements for different types of pattern recognition in conversational data (M. A. Walker, Passonneau, & Boland, 2001). Notably, this research highlights that more complex models do not consistently outperform simpler, smaller alternatives, emphasizing the need for theoretically informed model selection that accounts for domain-specific constraints rather than defaulting to complexity.

This research systematically evaluates models on real-world conversational datasets rather than synthetic or controlled data, contributing to AI theory (Deriu et al., 2021). It grounds AI evaluation in naturalistic, high-volume service interactions, resulting in empirical evidence that real-world data characteristics fundamentally influence model selection and performance. This suggests that theoretical advances in AI must explicitly account for context and domain specificity (Ding et al., 2018). The research demonstrates that domain-specific datasets can reveal insights about computational requirements and model effectiveness that are not visible when using standardized benchmarks, contributing to a more nuanced understanding of how AI systems perform across different real-world applications (Deriu et al., 2021).

The research further contributes to AI theory through advancing understanding

of how natural language processing and machine learning techniques can be applied to extract multiple types of information from conversational data from contact centers, from explicit content analysis to implicit emotional and behavioral pattern detection (Shah et al., 2023). The curiosity-driven learning framework, implemented through BDI agents, provides a theoretical foundation for autonomous, goal-directed knowledge exploration and extraction. This approach enables dynamic question generation, allowing AI systems to explore knowledge domains rather than passively processing fixed datasets. The framework supports adaptive information-seeking behavior for scalable knowledge extraction.

The CRISP-DM (Cross Industry Standard Process for Data Mining) framework guides a structured methodological approach that grounds these contributions (Martínez-Plumed et al., 2021). CRISP-DM enables a systematic alignment between domain-specific service challenges and AI model development by structuring the process into iterative phases of understanding, modeling, and evaluation. This structure not only ensures methodological transparency and reproducibility but also demonstrates how empirical insights from service data can guide theoretical advancements in model interpretability, multimodal learning, and human-in-the-loop systems. The cyclical nature of the framework supports theory building from empirical data. It promotes continuous refinement of models based on domain insights, reinforcing the importance of context-aware AI system development in real-world environments. This approach reinforces the importance of context-aware AI system development and its capacity to reveal new principles of machine learning performance in real-world applications, linking methodological rigor with theoretical insights.

8.4.3 Interdisciplinary contributions

These theoretical advances are directly related to the dissertation's broader interdisciplinary contributions, which bridge service management and computer science domains (Bardhan et al., 2010; Borges et al., 2021). The research provides methodologies for combining insights from technical and business domains, ensuring that technological advances are informed by and directly applicable to real service management challenges. The interdisciplinary approaches developed here demonstrate how findings can be relevant and applicable across different disciplinary perspectives.

This interdisciplinary approach demonstrates how cross-domain collaboration can

generate theoretical insights that neither field could achieve independently, advancing both service management and AI through their integration. Specifically, this dissertation shows that naturalistic service conversations contain learnable, structured, multi-layered patterns that can be systematically identified and utilized. The research makes these patterns explicit, providing a theoretical foundation for applying machine learning techniques to real-world service interactions and extending service management theory through demonstrating how multi-layered information flows can be operationalized for organizational learning, quality monitoring, and customer satisfaction analysis.

8.5 Practical contributions

8.5.1 Real-world data analysis

Beyond the theoretical contributions to artificial intelligence and service research, this dissertation provides substantial practical value for service organizations seeking to implement AI-augmented customer service systems. The practical relevance of these findings is supported by the use of real-world data from actual customer service operations rather than synthetic or laboratory-generated datasets (Deriu et al., 2021). The studies presented here analyzed real-life customer service interactions from naturalistic settings, including actual call center conversations, social media customer service exchanges, and authentic evaluations of service quality from operational service environments.

8.5.2 Actionable insights from individual studies

When implementing AI in customer service processes, organizations face critical decisions about signal identification, data requirements, and model selection (M. C. Lee, Scheepers, Liu, & Ngai, 2023). This dissertation addresses these implementation challenges through practical guidance. It demonstrates which signals can be effectively extracted from service interactions, identifies necessary data inputs for accurate analysis, and determines which model architectures perform optimally for various service applications. The research contributes actionable insights by systematically comparing model types to identify the most effective combinations with multiple data signals for specific service intelligence tasks (Ding et al., 2018).

The individual studies provide specific actionable insights for organizations implementing AI in their service operations. The systematic review of Chapter 2 demonstrates how technologies can augment, rather than replace, human service agents across various interaction stages, providing a framework for human-centered implementation. The speech emotion recognition research of Chapter 3 demonstrates that simple multi-layer perceptron models perform equivalently to complex neural networks for emotion detection, enabling cost-effective implementation choices. The customer dissatisfaction prediction study of Chapter 4 shows that audio processing capabilities are essential alongside text analysis, as combining multiple signals with cross-attention boosts prediction performance. The social media response strategy analysis in Chapter 5 reveals that automated multi-label classification can extract patterns of strategy effectiveness, with simple approaches outperforming expensive large language model solutions. Then, the quality monitoring study in Chapter 6 demonstrates that large language models can systematically identify errors from human evaluators, thereby creating hybrid review systems that enhance consistency while reducing manual workload. Finally, the BDI model performs general ontology extraction from document sets using dynamic, belief-based question modeling.

8.5.3 Strategic adoption of technology

More broadly, these findings enable organizations to create a competitive advantage through systematic extraction of previously hidden signals from service interactions (Choi, 2018). Organizations can analyze emotions, dissatisfaction, quality, and strategy effectiveness simultaneously to transform their processes from reactive problem-solving to proactive customer intelligence systems (Henkel, Bromuri, et al., 2020; Van Herck et al., 2022). However, successful implementation requires maintaining focus on human collaboration rather than replacement, ensuring technology enhances human capabilities while preserving the empathy and complex problem-solving skills that remain essential for service excellence (J. J. Kim et al., 2025).

The research provides organizations with both the “what” and the “how” of service intelligence extraction. The “what” encompasses systematic methodologies for converting previously unstructured service data into structured business insights across multiple dimensions simultaneously (Balducci & Marinova, 2018). Organizations gain confirmed approaches for processing large volumes of service interactions to extract actionable patterns that were previously invisible or accessible only through expen-

sive manual analysis (Bardhan et al., 2010). The “how” delivers practical model selection criteria with benchmarked performance comparisons across different modeling approaches. This combination of systematic information extraction capabilities and proven implementation enables organizations to transform service conversations into strategic assets, rather than treating them as merely operational costs.

The studies from this dissertation emphasize the strategic adoption of technology that complements existing organizational capabilities rather than replacing them. Organizations should prioritize implementing appropriate models, beginning with simple models before considering more complex alternatives. Future research directions could explore how training and development considerations for service agents shift from traditional quality monitoring to AI-augmented coaching, where human supervisors could use data-driven insights to provide service agents with more objective feedback on their interaction patterns and emotional regulation effectiveness. Such approaches enhance traditional supervisor-agent relationships through combining human mentoring with systematic analysis of measurable interaction outcomes.

8.6 Ethical and societal implications

The integration of AI into customer service processes, as explored in this dissertation, also comes with ethical and societal implications that extend beyond technical performance considerations. As these AI systems navigate the complex interplay between human relationships, organizational objectives, and technological capabilities, it results in various stakeholders who have different interests, concerns, and potential vulnerabilities (Wirtz et al., 2022). This research fundamentally involves real people, both as service agents who will work with these tools, as well as customers whose voices, emotions, and experiences become data input for algorithmic processing (Du & Xie, 2021). While this dissertation focuses primarily on the technical and theoretical aspects of AI implementation in service contexts, this multi-stakeholder reality necessitates careful examination of how these technologies affect human agency, privacy, fairness, and well-being across all parties involved in service delivery (Du & Xie, 2021).

8.6.1 Transformation of service work

With the introduction of AI in service processes, it has changed the way firms deal with customers (J. J. Kim et al., 2025). This necessitated a change in the skills of service agents, as they still require human expertise, but in combination with various technologies to assist them (Castaneda, Surachartkumtonkun, Maseeh, Thaichon, & Shao, 2025). However, this can lead to challenges such as skill dependency, potential erosion of intuitive judgment, and uncertainty about evolving professional roles (Bankins et al., 2023). These shifts in skills and expertise also shape team dynamics, influencing how human service agents collaborate with AI as a colleague and how they coordinate with other human agents in technologically mediated environments (de Jong, Schalk, & Curşeu, 2008). The continuous adaptation required to work with AI systems can create stress and questions about the balance between human decision-making and technological guidance in service interactions (Curşeu & Schruijer, 2012; Y. Huang & Gursoy, 2024).

The research presented in this dissertation, especially Chapter 2's literature review, positions technology as an augmentation tool instead of replacing service agents. These systems are not designed to perform autonomous decisions and actions, but only to aid the service agent in their work. Chapter 6, in particular, focuses on helping the service agents and firms improve service delivery, rather than automating performance evaluation. This augmentation approach preserves human agency while raising important questions about how service work evolves when service agents increasingly rely on AI-generated insights about customer emotions, conversation quality, and satisfaction predictors.

Another important consideration is how these AI systems affect the emotional aspects of service work (Bromuri et al., 2021). The technologies deployed in this dissertation, particularly those for emotion recognition and performance monitoring, alter how service agents approach emotional labor in their daily work. When AI can detect customer emotions, it creates a situation where service agents must balance their natural human empathy with the technology's insights. This raises concerns about whether service agents feel constantly monitored in their emotional responses and how they adjust their emotional labor when AI systems influence both customer interactions and their own performance (Holman, Chissick, & Totterdell, 2002). The integration of these technologies into emotional service interactions raises questions about the psychological well-being of service agents who must navigate between

authentic human connection and data-driven emotional insights, and warrants further investigation in future research.

8.6.2 Privacy, data protection, algorithmic fairness

The implementation of AI systems in service contexts invariably presents significant data-related challenges that require careful consideration. Service AI applications necessitate the collection and storage of various forms of personal data, which may raise concerns among stakeholders regarding privacy and data protection. Organizations must establish comprehensive policies and adhere to regulatory frameworks such as the General Data Protection Regulation (GDPR) to safeguard both customer and employee privacy rights (Sørum & Presthus, 2020). The AI systems examined in this dissertation process multiple data modalities, including voice signals, emotional expressions, and social media interactions, all of which contain sensitive personal information requiring stringent data governance protocols. While this research utilized secondary data that had undergone the companies' privacy review and compliance procedures, the broader deployment of such AI systems raises critical questions regarding data collection transparency, informed consent for emotion recognition technologies, and the ethical boundaries of organizational monitoring in service interactions involving both customers and employees (Iren et al., 2023).

Closely related to privacy and data protection is the topic of algorithmic fairness and potential bias in AI-driven service systems (Ntoutsi et al., 2020). This dissertation contributes to understanding these challenges, particularly through Chapter 6, where we examine the problems of AI systems in completing performance monitoring tasks. A key finding demonstrates that the data on which these AI systems are usually trained or benchmarked, specifically evaluations from human supervisors, contain inherent biases that can be perpetuated and amplified through algorithmic processing. Our research reveals that human evaluators themselves exhibit inconsistencies and biases in quality assessments, which means that AI systems trained on this data may institutionalize these subjective judgments rather than creating more objective evaluation processes. This contribution highlights how bias enters AI systems not just through algorithmic design, but fundamentally through the human-generated data used for training, emphasizing the need for organizations to address bias at the data collection and annotation level rather than assuming AI will eliminate human subjectivity (Celiktutan et al., 2024).

8.6.3 Energy consumption and environmental impact

The widespread deployment of AI systems for service information extraction raises important considerations regarding energy consumption and environmental sustainability (Budennyy et al., 2022). While this dissertation demonstrates the value of AI-driven extraction capabilities, the computational resources required for training and deploying these systems have significant energy implications. Large language models, in particular, require substantial computational power for both training and inference, contributing to carbon emissions and energy consumption (Iftikhar & Alsamhi, 2025).

In line with these concerns, our findings suggest that energy usage can be partially mitigated through the selection of appropriate models. Simpler or smaller models often achieve comparable performance to larger alternatives while consuming less computational power. This makes them a more energy-efficient option. Additionally, domain-specific adaptation shows that custom-trained models can outperform general-purpose large models, emphasizing that careful model selection not only supports performance but also contributes to more sustainable AI deployment in service contexts.

Organizations implementing AI-driven service analytics should consider energy-efficient model selection, balancing extraction capabilities with environmental impact (Budennyy et al., 2022). This includes evaluating whether the organizational learning benefits and service improvements justify the computational costs, and prioritizing efficient model architectures where performance allows. It will be important for future research to address sustainability in AI approaches for service information extraction that maintain analytical capabilities while minimizing environmental impact.

8.7 Limitations and avenues for future work

This dissertation has explored the application of machine learning and AI techniques to extract information from customer service conversations across multiple modalities and contexts. The research encompassed speech emotion recognition in call center interactions, multimodal customer satisfaction prediction from voice and text signals, automated detection of firm response strategies on social media, quality monitoring enhancement through large language models, and curiosity-driven knowledge extraction approaches. While these studies collectively demonstrate the poten-

tial for AI to augment human capabilities in service environments, they also reveal limitations that must be acknowledged and addressed in future research.

First, the studies in this dissertation utilized both open-access datasets and datasets from partner companies. While access to real-world company data represents a significant strength in terms of external validity, it also introduces some limitations (Deriu et al., 2021). Each study relied on a single, context-specific dataset, which allowed for in-depth exploration and practical relevance. While this enhances external validity, future work could explore the generalizability of these findings across more varied datasets and industries.

The proprietary nature of most datasets meant that research was constrained by what companies were willing to share, as organizations have concerns about privacy, competitive advantage, and regulatory compliance. Data collection was also typically confined to specific timeframes and limited organizational contexts, constraining our understanding of how findings might vary across different periods, companies, and industries. These limitations underscore the need for future research to focus on acquiring more diverse datasets across multiple companies, industries, and periods to enhance the generalizability and robustness of findings.

Second, although various types of models were implemented across the studies, these represent only a subset of the possible models available in the rapidly evolving AI domain. The fast growth of AI technologies in recent years has led to the emergence of new techniques at an unprecedented pace, creating both resource and time limitations in comprehensively evaluating all available approaches. Consequently, model selection was necessarily constrained by practical considerations rather than exhaustive comparison, though choices were made based on careful evaluation of available options at the time. The rapid technological evolution means that newer, potentially superior techniques may have emerged even during the course of this research. Therefore, our foundational implementations provide a baseline for future benchmarking, where future research should prioritize tracking emerging technologies and systematically evaluating how newer approaches perform in addressing these customer service problems, as significant improvements may be both possible and highly valuable for practical applications.

Next, these studies focused on extracting a specific set of signals from customer service conversations. However, many additional signals and outcomes could potentially be extracted and analyzed, such as brand-related outcomes (sales performance, customer churn, reputation metrics) or internal organizational indicators (em-

ployee stress levels, turnover rates, job satisfaction) (Katsikeas, Morgan, Leonidou, & Hult, 2016). Furthermore, the scope was limited to information extraction and analysis tools. At the same time, the broader domain of AI-augmented customer service includes other forms of technological support, such as personalized AI assistance systems, and emerging technologies in robotics, smart glasses, and other interactive devices that could fundamentally transform service delivery beyond the information extraction focus of this dissertation, as discussed in Chapter 2 (Bankins et al., 2023). Future research could therefore extend beyond the information extraction focus to examine these complementary technological developments.

Related to this expanded scope is the critical aspect of actual human-technology collaboration and the effects of these systems on all stakeholders involved. Customers may have concerns about being recorded or having their emotions automatically analyzed. At the same time, service agents must adapt to working with these technologies, potentially experiencing various psychological, social, and professional effects (Belanche et al., 2024). These human factors and stakeholder impacts were outside the scope of this dissertation but represent important areas for understanding the real-world implications of AI-augmented service systems (Bankins et al., 2023). Future research should investigate user acceptance across different stakeholder groups, organizational readiness for technology adoption, and the long-term effects of human-AI collaboration on both service providers and recipients (Walczuch, Lemmink, & Streukens, 2007).

Next, there is an ethical dimension to AI-augmented customer service that requires careful consideration as organizations navigate emerging regulations, such as the AI Act, and develop responsible deployment strategies (Wirtz et al., 2022). Ethical guidelines tailored explicitly to AI-augmented customer service contexts are essential to ensure that these systems are designed and deployed responsibly. Future research should focus on explainable AI approaches that enhance transparency and systematic investigation of biases in service AI systems. This includes ensuring that AI-enhanced services reduce rather than increase social inequalities. Additionally, research should prioritize creating adaptive research methodologies for quickly evaluating emerging technologies and building partnerships between academic researchers and technology developers. Finally, investigations into how AI augmentation can be designed to promote inclusive service delivery and support diverse cultural and individual needs remain critical (Du & Xie, 2021).

Finally, another promising avenue for future research involves the integration of

the signals discussed in this dissertation. Organizations can combine customer dissatisfaction identification with other forms of organizational knowledge, such as insights extracted from knowledge bases, ontologies, or conversational signals, developing more comprehensive and responsive service strategies (Marinova et al., 2016). This multi-layered approach would enable proactive problem resolution. Once dissatisfied customers are identified through the methods presented here, their concerns could be cross-referenced with existing organizational knowledge systems to understand root causes and identify patterns across different information sources. Subsequently, current service strategies could be systematically evaluated and updated based on these integrated insights, ultimately defining new service delivery paradigms that leverage learning from various organizational knowledge layers (Borges et al., 2021). This integration represents a significant step toward data-driven service innovation that goes beyond reactive problem-solving to encompass proactive strategy evolution based on comprehensive organizational intelligence (Choi, 2018).

In this work, we focused on implementing AI as an information extraction technology within customer service, demonstrating how such augmentation can improve efficiency and accuracy in handling customer requests. However, addressing the broader societal and organizational challenges that accompany AI-augmented services requires research that moves well beyond a purely technical approach. Future work will need to engage with multidisciplinary perspectives from psychology, ethics, law, sociology, and other relevant fields to understand how AI systems interact with human behavior, values, norms, and institutions. Such an approach is essential to ensure that AI systems in customer service evolve not only as powerful technical tools but also as responsible and inclusive components of service environments.

8.8 Concluding remarks

This dissertation demonstrated how AI can be utilized to extract meaningful information from real-world customer service interactions. Across several studies, it was shown that machine learning models can process everyday service conversations to uncover patterns that support better-informed decision-making, service evaluation, and process improvements. More fundamentally, this research bridges a critical gap between the rich information potential embedded in service encounters and orga-

nizations' ability to systematically extract and utilize this knowledge, contributing to both service management and information systems theory by establishing AI-driven extraction capabilities as organizational resources.

Throughout the chapters, different data modalities, outcome measures, and modeling techniques were explored. This shows that real-life service interactions are complex and require flexible methods to capture their various dimensions of multi-layered information. It demonstrates that content, implicit, and aggregated layers of service interactions can be analyzed through diverse AI approaches. The systematic comparison of AI approaches across service contexts reveals that complex models do not consistently outperform simpler alternatives. At the same time, domain-specific adaptation often surpasses general-purpose approaches, contributing methodological foundations for context-appropriate AI selection in service management.

The findings demonstrate that AI can extract valuable information from customer service conversations. This research establishes that customer-firm interactions are essential resources containing rich patterns that can be systematically processed and that tacit knowledge embedded in service interactions can be transformed into explicit organizational assets. These advances contribute to service management theory by extending service-dominant logic through systematic methods for analyzing information exchange processes, and to organizational learning theory by demonstrating how AI-driven extraction enhances organizational information processing capabilities.

Together, these findings contribute to a more informed and evidence-based approach to using AI in service environments. This dissertation provides both theoretical insights into how service interactions can be systematically analyzed as multi-layered information sources and practical guidance for developing future systems that can extract valuable intelligence from customer service conversations, establishing these encounters as repositories of organizational knowledge that can enhance decision-making and organizational learning processes.

References

Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, *10*, 122136–122158. doi: 10.1109/ACCESS.2022.3223444

Abi Kanaan, M., Couchot, J.-F., Guyeux, C., Laiymani, D., Atechian, T., & Darazi, R. (2024). Combining a multi-feature neural network with multi-task learning for emergency calls severity prediction. *Array*, *21*, 1000333. doi: 10.1016/j.array.2023.100333

Agichtein, E., & Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries* (pp. 85–94). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/336597.33664

Ahmed, A., Shaalan, K., Toral, S., & Hifny, Y. (2021). A multimodal approach to improve performance evaluation of call center agent. *Sensors*, *21*(8), 2720–2731. doi: 10.3390/s21082720

Ahmed, A., Sivarajah, U., Irani, Z., Mahroof, K., & Charles, V. (2024). Data-driven subjective performance evaluation: An attentive deep neural networks model based on a call centre case. *Annals of Operations Research*, *333*, 939–970. doi: 10.1007/s10479-022-04874-2

Ahmed, A., Toral, S., Shaalan, K., & Hifny, Y. (2020). Agent productivity modeling in a call center domain using attentive convolutional neural networks. *Sensors*, *20*(19), 5489–5500. doi: 10.3390/s20195489

Ahmed, C., ElKorany, A., & ElSayed, E. (2023). Prediction of customer's perception in social networks by integrating sentiment analysis and machine learning. *Journal of Intelligent Information Systems*, *60*, 829–851. doi: 10.1007/s10844-022-00756-y

Aksin, Z., Armony, M., & Mehrotra, V. (2009). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, *16*(6), 665–688. doi: 10.1111/j.1937-5956.2007.tb00288.x

Aksoy, L., Buoye, A., Aksoy, P., Larivière, B., & Keiningham, T. L. (2013). A cross-national investigation of the satisfaction and loyalty linkage for mobile telecommunications services across eight countries. *Journal of Interactive Marketing*, *27*(1), 74–82. doi: 10.1016/j.intmar .2012.09.003

Alaimo, C., & Kallinikos, J. (2021). Organizations decentered: Data objects, technology and knowledge. *Organization Science*, *33*(1). doi: 10.1287/orsc.2021.1552

Alam, F., Danieli, M., & Riccardi, G. (2018). Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, *50*, 40–60. doi: 10.1016/j.csl.2017.12.003

Al-Aswadi, F. N., Chan, H. Y., & Gan, K. H. (2020). Automatic ontology construction from text: A review from shallow to deep learning trend. *Artificial Intelligence Review*, *53*, 3901–3928. doi: 10.1007/s10462-019-09782-9

Albrecht, T., Rausch, T. M., & Derra, N. D. (2021). Call me maybe: Methods and practical implementation of artificial intelligence in call center arrivals' forecasting. *Journal of Business Research*, *123*, 267–278. doi: 10.1016/j.jbusres.2020.09.033

Ali Zaidi, S. S., Fraz, M. M., Shahzad, M., & Khan, S. (2021). A multiapproach generalized framework for automated solution suggestion of support tickets. *International Journal of Intelligent Systems*, *37*(6), 3654–3681. doi: 10.1002/int.22701

Al-Mutawa, R. F., & Al-Aama, A. Y. (2024). Arabic opinion classification of customer service conversations using data augmentation and artificial intelligence. *Big Data and Cognitive Computing*, *8*(12), 196–217. doi: 10.3390/bdcc8120196

Ameen, N., Tarhini, A., Reppel, A., & Anand, A. (2011). Customer experiences in the age of artificial intelligence. *Computers in Human Behavior*, *114*, 106548. doi: 10.1016/j.chb.2020 .106548

An, S., Ma, Z., Lin, Z., Zheng, N., Lou, J.-G., & Chen, W. (2024). Make your LLM fully utilize the context. In A. Globerson et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 37, pp. 62160–62188).

Ando, A., Masumura, R., Kamiyama, H., Kobashikawa, S., & Aono, Y. (2017). Hierarchical LSTMs with joint learning for estimating customer satisfaction from contact center calls. In *Interspeech.*

Ando, A., Masumura, R., Kamiyama, H., Kobashikawa, S., Aono, Y., & Toda, T. (2020). Customer satisfaction estimation in contact center calls based on a hierarchical multi-task model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 715–728. doi: 10.1109/TASLP.2020.2966857

Andreassen, T. W. (1999). What drives customer loyalty with complaint resolution? *Journal of Service Reserach*, *1*(4), 324–332. doi: 10.1177/109467059914004

Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-Net: A lightweight CNN-based speech emotion recognition system using deep frequency features. *Sensors*, *20*(18), 5212. doi: 10.3390/s20185212

Arwin, H., Halldórsson, Á., & Hellström, A. (2024). Advancing relational primary healthcare: Four triadic components of the digital face-to-face professional service encounter. *European Management Journal*. doi: 10.1016/j.emj.2024.11.009

Ashtar, S., Yom-Tov, G. B., Akiva, N., & Rafaeli, A. (2021). When do service employees smile? Response-dependent emotion regulation in emotional labor. *Journal of Organizational Behavior*, *42*(9), 1202–1227. doi: 10.1002/job.2562

Ashtar, S., Yom-Tov, G. B., Rafaeli, A., & Wirtz, J. (2023). Affect-as-information: Customer and employee affective displays as expeditious predictors of customer satisfaction. *Journal of Service Research*, *27*(4), 525–542. doi: 10.1177/10946705231194076

Azemi, Y., Ozuem, W., & Howell, K. E. (2020). The effects of online negative word-of-mouth on dissatisfied customers: A frustration-aggression perspective. *Psychology & Marketing*, *37*(4), 564–577. doi: 10.1002/mar.21326

Babaei Giglou, H., D'Souza, J., & Auer, S. (2023). LLMs4OL: Large language models for ontology learning. In T. R. Payne et al. (Eds.), *The Semantic Web – ISWC 2023* (pp. 408–427). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-47240-4_22

Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Young Lee, M., … Baik, S. W. (2019). Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*, *78*, 5571–5589. doi: 10.1007/s11042-017-5292-7

Baier, L., Kühl, N., Schüritz, R., & Satzger, G. (2021). Will the customers be happy? Identifying unsatisfied customers from service encounter data. *Journal of Service Management*, *32*(2), 265–288. doi: 10.1108/JOSM-06-2019-0173

Bailey, C., & Clark, M. (2007, September). *How companies use customer insight in inbound service call centres to drive cross-selling, up-selling and retention: An exploratory multiple case study* (Research Report No. R10). Henley Centre for Customer Management. Retrieved from `https://centaur.reading.ac.uk/83733/1/R10_Customer%20Insight%20inbound%20call%20centres%20(Sept%202007).pdf` (Accessed: 16 January 2025)

Balaji, T., Annavarapu, C. S. R., & Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, *40*, 100395. doi: 10.1016/j.cosrev.2021.100395

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*, 557–590. doi: 10.1007/s11747-018-0581-x

Bankins, S., Ocampo, A. C., Marrone, M., Restubog, S. L. D., & Woo, S. E. (2023). A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *Journal of Organizational Behavior*, *45*(2), 159–182. doi: 10.1002/job.2735

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636. doi: 10.1037/0022-3514.70.3.614

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). *TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification.* Retrieved from `https://arxiv.org/abs/2010.12421`

Bardhan, I. R., Demirkan, H., Kannan, P., Kauffman, R. J., & Sougstad, R. (2010). An interdisciplinary perspective on IT services management and service science. *Journal of Management Information Systems*, *26*(4), 13–64. doi: 10.2753/MIS0742-1222260402

Barrett, M., Oborn, E., & Orlikowski, W. (2016). Creating value in online communities: The sociomaterial configuring of strategy, platform, and stakeholder engagement. *Information Systems Research*, *27*(4), 704–723. doi: 10.1287/isre.2016.0648

Bassili, J. N. (1979). Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, *37*(11), 2049–2058. doi: 10.1037/0022-3514.37.11.2049

Belanche, D., Belk, R. W., Casaló, L. V., & Flavián, C. (2024). The dark side of artificial intelligence in services. *The Service Industries Journal*, *44*(3-4), 149–172. doi: 10.1080/02642069.2024.2305451

Benayas, A., Sicilia, M. A., & Mora-Cantallops, M. (2024). A comparative analysis of encoder only and decoder only models in intent classification and sentiment analysis: Navigating the trade-offs in model size and performance. *Language Resources & Evaluation*. doi: 10.1007/s10579-024-09796-y

Blaurock, M., Büttgen, M., & Schepers, J. (2024). Designing collaborative intelligence systems for employee-AI service co-production. *Journal of Service Research*, *0*(0), 1–19. doi: 10.1177/10946705241238751

Blaurock, M., Čaić, M., Okan, M., & Henkel, A. P. (2022). A transdisciplinary review and framework of consumer interactions with embodied social robots: Design, delegate, and deploy. *International Journal of Consumer Studies*, *46*(5), 1877–1899. doi: 10.1111/ijcs.12808

Blaurock, v. M., Marah, Okan, M., & Henkel, A. P. (2022). Robotic role theory: An integrative review of human-robot service interaction to advance role theory in the age of social robots. *Journal of Service Management*, *33*(6), 27–49. doi: 10.1108/JOSM-09-2021-0345

Bloemer, J., de Ruyter, K., & Wetzels, M. (1999). Linking perceived service quality and service loyalty: A multi-dimensional perspective. *European Journal of Marketing*, *33*(11/12), 1082–1106. doi: 10.1108/03090569910292285

Bockhorst, J., Yu, S., Polania, L., & Fung, G. (2017). Predicting self-reported customer satisfaction of interactions with a corporate call center. In *Machine Learning and Knowledge Discovery in Databases* (pp. 179–190). Springer International Publishing. doi: 10.1007/978-3-319-71273-4_15

Bohne, R. (2024). *Communication channels customers prefer to use to resolve customer service issues in the United States in 2022.* `https://www.statista.com/statistics/818566/preferred-channels-customer-service-issues-united-states/`. Statista Research Department. (Accessed: 2024-10-09)

Bohne, Raphael. (2022). *What is your expected response time for social media questions or complaints.* `https://www.statista.com/statistics/808477/expected-response-time-for-social-media-questions-or-complaints/`. Statista Research Department. (Accessed: 2024-10-09)

Bohne, Raphael. (2024). *Number of contact center employees in the united states from 2014 to 2023.* `https://www.statista.com/statistics/881114/contact-center-employees-united-states`. Statista Research Department. (Accessed: 2025-01-16)

Bojanić, M., Delić, V., & Karpov, A. (2020). Call redistribution for a call center based on speech emotion recognition. *Applied Sciences*, *10*(13), 4653. doi: 10.3390/app10134653

Borah, S. B., Prakhya, S., & Sharma, A. (2020). Leveraging service recovery strategies to reduce customer churn in an emerging market. *Journal of the Academy of Marketing Science*, *48*, 848–868. doi: 10.1007/s11747-019-00634-0

Borg, A., Boldt, M., Rosander, O., & Ahlstrand, J. (2021). E-mail classification with machine learning and word embeddings for improved customer support. *Neural Computing and Applications*, *33*, 1881–1902. doi: 10.1007/s00521-020-05058-4

Borges, A. F., Laurindo, F. J., Spínola, M., Mauro, Gonçalves, R. F., & Mattos, C. A. (2021). The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions. *International Journal of Information Management*, *57*, 102225. doi: 10.1016/j.ijinfomgt.2020.102225

Bost, X., Senay, G., El-Bèze, M., & De Mori, R. (2015). Multiple topic identification in human/human conversations. *Computer Speech & Language*, *34*(1), 18–42. doi: 10.1016/j.csl.2015.03.006

Bougie, R., Pieters, R., & Zeelenberg, M. (2003). Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. *Journal of the Academy of Marketing Sciences*, *31*(4), 377–393. doi: 10.1177/009207030325441

Bowen, D. E. (2016). The changing role of employees in service theory and practice: An interdisciplinary view. *Human Resource Management Review*, *26*(1), 4–13. doi: 10.1016/j.hrmr.2015.09.002

Bowen, D. E. (2024). An organizational behavior/human resource management perspective on the roles of people in a service organization context: Frameworks and themes. *Journal of Service Management*, *35*(1), 1–21. doi: 10.1108/JOSM-10-2023-0424

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. doi: 10.1016/S0031-3203(96)00142-2

Bratman, M. E. (1987). *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Breit, E., Egeland, C., Løberg, I. B., & Røhnebæk, M. T. (2020). Digital coping: How frontline workers cope with digital service encounters. *Social Policy & Administration*, *55*(5), 833–847. doi: 10.1111/spol.12664

Breuer, K., Nieken, P., & Sliwka, D. (2013). Social ties and subjective performance evaluations: An empirical investigation. *Review of Managerial Science*, *7*, 141–157. doi: 10.1007/s11846-011-0076-3

Bromuri, S., Henkel, A. P., Iren, D., & Urovi, V. (2021). Using AI to predict service agent stress from emotion patterns in service interactions. *Journal of Service Management*, *32*(4), 581–611. doi: 10.1108/JOSM-06-2019-0163

Bruni, R., Bianchi, G., & Papa, P. (2023). Hyperparameter black-box optimization to improve the automatic classification of support tickets. *Algorithms*, *16*(1), 46–60. doi: 10.3390/a16010046

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, *140*(2), 889–942. doi: 10.1093/qje/qjae044

Budennyy, S. A., Lazarev, V. D., Zakharenko, N. N., Korovin, A. N., Plosskaya, O. A., Dimitrov, D. V., … Zhukov, L. E. (2022). eco2AI: Carbon emissions tracking of machine learning models as the first step towards sustainable AI. *Doklady Mathematics*, *106*, 118–128. doi: 10.1134/S1064562422060230

Burgoon, J. K., Bonito, J. A., Ramirez Jr., A., Dunbar, N. E., Kam, K., & Fischer, J. (2006). Testing the interactivity principle: Effects of mediation, propinquity, and verbal and nonverbal modalities in interpersonal interaction. *Journal of Communication*, *52*(3), 657–677. doi: 10 .1111/j.1460-2466.2002.tb02567.x

Büyük, O. (2024). A comprehensive evaluation of large language models for Turkish abstractive dialogue summarization. *IEEE Access*, *12*, 124391–124401. doi: 10.1109/ ACCESS.2024.3454342

Cai, N., Li, S., Xu, J., Tian, Y., Zhou, Y., & Liao, J. (2025). A hybrid intention recognition framework with semantic inference for financial customer service. *Electronics*, *14*(3), 495–515. doi: 10.3390/electronics14030495

Cambra-Fierro, J., Melero, I., & Sese, F. J. (2015). Managing complaints to improve customer profitability. *Journal of Retailing*, *91*(1), 109–124. doi: 10.1016/j.jretai.2014.09.004

Carroll, N., & Conboy, K. (2020). Normalising the “new normal”: Changing tech-driven work practices under pandemic time pressure. *International Journal of Information Management*, *55*, 102186. doi: 10.1016/j.ijinfomgt.2020.102186

Caruana, A. (2002). Service loyalty: The effects of service quality and the mediating role of customer satisfaction. *European Journal of Marketing*, *36*(7/8), 811–828. doi: 10.1108/ 03090560210430818

Carvalho, I., Oliveira, H. G., & Silva, C. (2023). The importance of context for sentiment analysis in dialogues. *IEEE Access*, *11*, 86088–86103. doi: 10.1109/ACCESS.2023.3304633

Castaneda, A. R., Surachartkumtonkun, J., Maseeh, H. I., Thaichon, P., & Shao, W. (2025). Frontline employees in an AI-integrated workplace: Current perspectives and future research landscapes. *Journal of Service Theory*, *35*(7), 30–60. doi: 10.1108/JSTP-04-2024-0099

Castelo, N., Boegershausen, J., Hildebrand, C., & Henkel, A. P. (2023). Understanding and improving consumer reactions to service bots. *Journal of Consumer Research*, *50*(4), 848–863. doi: 10.1093/jcr/ucad023

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. doi: 10.1177/0022243719851788

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, *2017*(108), 109–120. doi: 10.1515/pralin-2017-0013

Celiktutan, B., Cadario, R., & Morewedge, C. K. (2024). People see more of their biases in algorithms. *Psychological and Cognitive Sciences*, *121*(16), e2317602121. doi: 10.1073/pnas.2317602121

Chacón, H., Koppisetti, V., Hardage, D., Choo, K.-K. R., & Rad, P. (2023). Forecasting call center arrivals using temporal memory networks and gradient boosting algorithm. *Expert Systems with Applications*, *224*, 119983. doi: 10.1016/j.eswa.2023.119983

Chang, C.-C., & Hung, J.-S. (2018). The effects of service recovery and relational selling behavior on trust, satisfaction, and loyalty. *International Journal of Bank Marketing*, *36*(7), 1437–1454. doi: 10.1108/IJBM-07-2017-0160

Cheang, H. S., & Pell, M. D. (2008). The sound of sarcasm. *Speech Communication*, *50*(5), 366–381. doi: 10.1016/j.specom.2007.11.003

Chen, D., Zhengwei, H., Jintao, M., & Khanal, R. (2024). Sentiments analysis for intelligent customer service dialogue using hybrid word embedding and stacking ensemble. *Soft Computing*, *28*, 11619–11631. doi: 10.1007/s00500-024-09899-2

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), 1165–1188. doi: 10.2307/41703503

Chi, O. H., Denton, G., & Gursoy, D. (2020). Artificially intelligent device use in service delivery: A systematic review, synthesis, and research agenda. *Journal of Hospitality Marketing & Management*, *29*(7), 757–786. doi: 10.1080/19368623.2020.1721394

Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: A systematic review of deep learning and machine learning approaches. *Journal of Big Data*, *9*(10). doi: 10.1186/s40537-022-00561-y

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In D. Wu, M. Carpuat, X. Carreras, &

E. M. Vecchi (Eds.), *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103–111). Doha, Qatar: Association for Computational Linguistics. doi: 10.3115/v1/w14-4012

Choi, S. (2018). Organizational knowledge and information technology: The key resources for improving customer service in call centers. *Information Systems and e-Business Management*, *16*, 187–203. doi: 10.1007/s10257-017-0359-6

Choi, S., & Kim, S. (2025). Consumer perception of employees with disabilities using robots. *Annals of Tourism Research*, *112*, 103945. doi: 10.1016/j.annals.2025.103945

Cimiano, P., Hotho, A., & Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, *24*, 305–339. doi: 10.1613/jair.1648

Clarke, V., & Braun, V. (2014). Thematic analysis. In T. Teo (Ed.), *Encyclopedia of Critical Psychology.* New York, NY: Springer New York.

Cocarascu, O., & Toni, F. (2018). Combining deep learning and argumentative reasoning for the analysis of social media textual content using small data sets. *Computational Linguistics*, *44*(4), 833–858. doi: 10.1162/coli_a_00338

Cohen, M. C. (2018). Big data and service operations. *Production and Operations Management*, *27*(9), 1709–1723. doi: 10.1111/poms.1283

Cong, P., Wang, C., Ren, Z., Wang, H., Wang, Y., & Feng, J. (2016). Unsatisfied customer call detection with deep learning. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 1–5). doi: 10.1109/ISCSLP.2016.7918385

Cortiz, D. (2021). *Exploring Transformers in Emotion Recognition: a comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA.* Retrieved from `https://arxiv.org/abs/2104.02041`

Crivellari, A., & Beinat, E. (2020). Trace2trace- A feasibility study on neural machine translation applied to human motion trajectories. *Sensors*, *20*(12), 3503. doi: 10.3390/s20123503

Curşeu, P. L., & Schruijer, S. G. L. (2012). Decision styles and rationality: An analysis of the predictive validity of the general decision-making style inventory. *Educational and Psychological Measurement*, *72*(6), 1053–1062. doi: 10.1177/0013164412448066

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, *32*(5), 554–571. doi: 10.1287/mnsc.32.5.554

Dangol, R., Alsadoon, A., Prasad, P., Seher, I., & Alsadoon, O. H. (2020). Speech emotion recognition using convolutional neural network and long-short term memory. *Multimedia Tools and Applications*, *79*, 32917–32934. doi: 10.1007/s11042-020-09693-w

Davies, R., Coole, T., & Smith, A. (2017). Review of socio-technical considerations to ensure successful implementation of industry 4.0. *Procedia Manufacturing*, *11*, 1288–1295. doi: 10.1016/j.promfg.2017.07.256

Davoust, A., Gavigan, P., Ruiz-Martin, C., Trabes, G., Esfandiari, B., Wainer, G., & James, J. (2020). An architecture for integrating BDI agents with a simulation environment. In *Engineering Multi-Agent Systems* (pp. 67–84). doi: 10.1007/978-3-030-51417-4_4

DeBellis, M., Dutta, N., Gino, J., & Balaji, A. (2024). Integrating ontologies and large language models to implement retrieval augmented generation. *Applied Ontology*, *19*(4), 389–407. doi: 10.1177/15705838241296446

De Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, J. K.-U., & Von Wangenheim, F. (2022). Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing*, *51*(1), 91–105. doi: 10.1016/j.intmar.2020.04.007

De Cleen, T., Baecke, P., & Goedertier, F. (2025). The influence of emotions and communication style on customer satisfaction and recommendation in a call center context: An NLP-based analysis. *Journal of Business Research*, *189*, 115192. doi: 10.1016/j.jbusres.2025.115192

De Gauquier, L., Willems, K., Cao, H.-L., Vanderborght, B., & Brengman, M. (2023). Together or alone: Should service robots and frontline employees collaborate in retail-customer interactions at the POS? *Journal of Retailing and Consumer Services*, *70*, 103176. doi: 10.1016/j.jretconser.2022.103176

de Jong, R., Schalk, R., & Curşeu, P. L. (2008). Virtual communicating, conflicts and performance in teams. *Team Performance Management*, *14*(7/8), 364–380. doi: 10.1108/13527590810912331

De Keyser, A., Köcher, S., Alkire, L. n. N., Verbeeck, C., & Kandampully, J. (2019). Frontline service technology infusion: Conceptual archetypes and future research directions. *Journal of Service Management*, *30*(1), 156–183. doi: 10.1108/JOSM-03-2018-0082

de Lacerda Pataca, C., & Costa, P. D. P. (2023). Hidden bawls, whispers, and yelps: Can text convey the sound of speech, beyond words? *IEEE Transactions on Affective Computing*, *14*(1), 6–16. doi: 10.1109/TAFFC.2022.3174721

Deniz, E., Erbay, H., & Coşar, M. (2022). Multi-label classification of e-commerce customer reviews via machine learning. *Axioms*, *11*(9), 436. doi: 10.3390/axioms11090436

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, *54*, 755–810. doi: 10.1007/s10462-020-09866-x

De Ruyter, K., Wetzels, M., & Feinberg, R. (2001). Role stress in call centers: Its effects on employee performance and satisfaction. *Journal of Interactive Marketing*, *15*(2), 23–35. doi: 10.1002/dir.1008

de Velasco, M., Justo, R., Antón, J., Carrilero, M., & Torres, M. I. (2018). Emotion detection from speech and text. In *IberSPEECH* (pp. 68–71). doi: 10.21437/iberspeech.2018-15

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 0, pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-1423

Dhiman, N., & Kumar, A. (2022). What we know and don't know about consumer happiness: Three-decade review, synthesis, and research propositions. *Journal of Interactive Marketing*, *58*(2-3), 115–135. doi: 10.1177/10949968221095548

Ding, J., Tarokh, V., & Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, *35*(6), 16–34. doi: 10.1109/MSP.2018.2867638

Dobrucali Yelkenci, B., Özdağoğlu, G., & İlter, B. (2023). Online complaint handling: A text analytics-based classification framework. *Marketing Intelligence & Planning*, *41*(5), 557–573. doi: 10.1108/MIP-05-2022-0188

Dolata, M., Agotai, D., Schubiger, S., & Schwabe, G. (2020). Advisory service support that works: Enhancing service quality with a mixed-reality system. *Proc. ACM Hum.-Comput. Interact.*, *4*(CSCW2). doi: 10.1145/3415191

Dong, C.-S. J., & Srinivasan, A. (2013). Agent-enabled service-oriented decision support systems. *Decision Support Systems*, *55*(1), 264–373. doi: 10.1016/j.dss.2012.05.047

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., … Jégou, H. (2025). *The Faiss library.* Retrieved from `https://arxiv.org/abs/2401.08281`

Du, S., & Xie, C. (2021). Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities. *Journal of Business Research*, *129*, 961–974. doi: 10.1016/j.jbusres.2020.08.024

Duarte, F. (2024). *X (formerly Twitter) user age, gender, & demographic stats.* https://explodingtopics.com/blog/x-user-stats. (Accessed: 2024-10-09)

Dukes, A., & Zhu, Y. (2019). Why customer service frustrates consumers: Using a tiered organizational structure to exploit hassle costs. *Marketing Science*, *38*(3), 500–515. doi: 10.1287/mksc.2019.1149

Ebadi Jalal, M., Hosseini, M., & Karlsson, S. (2016). Forecasting incoming call volumes in call centers with recurrent neural networks. *Journal of Business Research*, *69*(11), 4811–4814. doi: 10.1016/j.jbusres.2016.04.035

Edvardsson, B., Tronvoll, B., & Gruber, T. (2011). Expanding understanding of service exchange and value co-creation: A social construction approach. *Journal of the Academy of Marketing Science*, *39*, 327–339. doi: 10.1007/s11747-010-0200-y

Einwiller, S. A., & Steilen, S. (2015). Handling complaints on social network sites - An analysis of complaints and complaint responses on Facebook and Twitter pages of large US companies. *Public Relations Review*, *41*(2), 195–204. doi: 10.1016/j.pubrev.2014.11.012

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3-4), 169–200. doi: 10.1080/02699939208411068

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., … Ricci-Bitti, P. E. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, *53*(4), 712–717. doi: 10.1037/0022-3514.53.4.712

Elnaga, A., & Imran, A. (2013). The effect of training on employee performance. *European Journal of Business and Management*, *5*(4), 137–147.

Falagas, E., Matthew, Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2007). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal*, *22*(2), 338–342. doi: 10.1096/fj.07-9492LSF

Fan, A., & Mattila, A. S. (2020). Touch versus tech in service encounters. *Cornell Hospitality Quarterly*, *62*(4), 468–481. doi: 10.1177/1938965520957282

Fan, S., & Ilk, N. (2020). A text analytics framework for automated communication pattern analysis. *Information & Management*, *57*(4), 103219. doi: 10.1016/j.im.2019.103219

Fernández-Sabiote, E., & López-López, I. (2020). Discovering call interaction fluency: A way to improve experiences with call centres. *Service Science*, *12*(1), 26–42. doi: 10.1287/serv.2019.0251

Folan, P., & Browne, J. (2005). A review of performance management: Towards performance management. *Computers in Industry*, *56*(7), 663–680. doi: 10.1016/j.compind.2005.03.001

Forbes, & Ravinutala, S. (2023). *From emotion to empathy: Bringing human experience to voice ai.* Forbes Technology Council. Retrieved from `https://www.forbes.com/sites/forbestechcouncil/2022/09/22/from-emotion-to-empathy-bringing-human-experience-to-voice-ai/` (Accessed: September 22, 2023)

Fraering, M., & Minor, M. S. (2013). Beyond loyalty: Customer satisfaction, loyalty, and fortitude. *Journal of Services Marketing*, *27*(4), 334–344. doi: 10.1108/08876041311330807

Frering, L., Steinbauer-Wagner, G., & Holzinger, A. (2025). Integrating belief-desire-intention agents with large language models for reliable human-robot interaction and explainable artificial intelligence. *Engineering Applications of Artificial Intelligence*, *141*, 109771. doi: 10.1016/j.engappai.2024.109771

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, *97*(3), 412–429. doi: 10.1037/0033-2909.97.3.412

Frijda, N. H., Mesquita, B., Sonnemans, J., & van Goozen, S. H. M. (1991). The duration of affective phenomena or emotions, sentiments and passions. In K. T. Strongman (Ed.), *International review of studies on emotion* (Vol. 1, pp. 187–225).

Fuentes, M., Smyth, H., & Davies, A. (2019). Co-creation of value outcomes: A client perspective on service provision in projects. *International Journal of Project Management*, *37*(5), 695–715. doi: 10.1016/j.ijproman.2019.01.003

Galal, M. A., Yousef, A. H., Zayed, H. H., & Medhat, W. (2024). Arabic sarcasm detection: An enhanced fine-tuned language model approach. *Ain Shams Engineering Journal*, *15*(6), 102736. doi: 10.1016/j.asej.2024.102736

Gao, W., Fan, H., Li, W., & Wang, H. (2021). Crafting the customer experience in omnichannel contexts: The role of channel integration. *Journal of Business Research*, *126*, 12–22. doi: 10.1016/j.jbusres.2020.12.056

Garry, T., & Harwood, T. (2019). Cyborgs as frontline service employees: A research agenda. *Journal of Service Theory and Practice*, *29*(4), 415–437. doi: 10.1108/JSTP-11-2018-0241

Gavigan, P., & Esfandiari, B. (2022). BDI for autonomous mobile robot navigation. In *Engineering Multi-Agent Systems* (Vol. 13190, pp. 137–155). Cham: Springer International Publishing. doi: 10.1007/978-3-030-97457-2_8

Gemma Team, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., … Andreev, A. (2024). *Gemma 2: Improving open language models at a practical size.* Retrieved from https://arxiv.org/abs/2408.00118

George, A. S., Baskar, T., & Srikaanth, P. B. (2024). The erosion of cognitive skills in the technological age: How reliance on technology impacts critical thinking, problem-solving, and creativity. *Partners Universal Innovative Research Publication*, *2*(3), 147–163. doi: 10.5281/zenodo.11671150

Gheini, M., Ren, X., & May, J. (2021). *On the strengths of cross-attention in pretrained transformers for machine translation.* Retrieved from https://arxiv.org/abs/2104.08771 (Accessed: 16 January 2025)

Giannakopoulos, T. (2015). pyAudioAnalysis: An open-source Python library for audio signal analysis. *PLOS One*, *10*(12). doi: 10.1371/journal.pone.0144610

Giovanna, N., & Luciana, D. V. (2011). Errors in customer satisfaction surveys and methods to correct self-selection bias. *Quality Technology & Quantitative Management*, *8*(2), 167–181. doi: 10.1080/16843703.2011.11673254

Gnewuch, U., Morana, S., Hinz, O., Kellner, R., & Maedche, A. (2023). More than a bot? The impact of disclosing human involvement on customer interactions with hybrid service agents. *Information Systems Research*, *35*(3), 936–955. doi: 10.1287/isre.2022.0152

Goffin, K. (1999). Customer support: A cross-industry study of distribution channels and strategies. *International Journal of Physical Distribution & Logistics Management*, *29*(6), 374–398. doi: 10.1108/09600039910283604

Goldberg, L. S., & Grandey, A. A. (2007). Display rules versus display autonomy: Emotion regulation, emotional exhaustion, and task performance in a call center simulation. *Journal of Occupational Health Psychology*, *12*(3), 301–318. doi: 10.1037/1076-8998.12.3.301

González-Docasal, A., Pérez, N., Alvarez, A., Serras, M., García-Sardiña, L., Arzelus, H., … Romero, B. (2020). Nalytics: Natural speech and text analytics. *Natural Language Processing*, *65*, 119–122. doi: 10.26342/2020-65-17

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, *4*(3), 379–391. doi: 10.1016/j.joi.2010.03.002

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly*, *19*(2), 213–236. doi: 10.2307/249689

Grandey, A. A., Dickter, D. N., & Sin, H.-P. (2004). The customer is not always right: Customer aggression and emotion regulation of service employees. *Journal of Organizational Behavior*, *25*(3), 397–418. doi: 10.1002/job.252

Grewal, D., Herhausen, D., Ludwig, S., & Ordenes, F., Villarroel Ordenes. (2022). The future of digital communication research: Considering dynamics and multimodality. *Journal of Retailing*, *98*(2), 224–240. doi: 10.1016/j.jretai.2021.01.007

Grewal, D., Kroschke, M., Mende, M., Roggeveen, A. L., & Scott, M. L. (2022). Frontline cyborgs at your service: How human enhancement technologies affect customer experiences in retail, sales, and service settings. *Journal of Interactive Marketing*, *51*(1), 9–25. doi: 10.1016/j.intmar.2020.03.001

Grljević, O., & Bošnjak, Z. (2018). Sentiment analysis of customer data. *Strategic Management*, *23*(3), 38–49. doi: 10.5937/StratMan1803038G

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. doi: 10.1006/knac.1993.1008

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, *43*(5-6), 907–928. doi: 10.1006/ijhc.1995.1081

Grönroos, C. (2011). Value co-creation in service logic: A critical analysis. *Marketing Theory*, *11*(3), 279–301. doi: 10.1177/1470593111408177

Guo, Y., Fan, D., & Zhang, X. (2020). Social media-based customer service and firm reputation. *International Journal of Operations & Production Management*, *40*(5), 575–601. doi: 10.1108/IJOPM-04-2019-0315

Guo, Y., Li, Y., Liu, D., & Xu, S. X. (2024). Measuring service quality based on customer emotion: An explainable AI approach. *Decision Support Systems*, *176*, 114051. doi: 10.1016/j.dss.2023.114051

Hadifar, A., Labat, S., Hoste, V., Develder, C., & Demeester, T. (2021). *A million tweets are worth a few points: Tuning transformers for customer service tasks.* Retrieved from `https://arxiv.org/abs/2104.07944`

Haenlein, M., & Kaplan, A. M. (2012). The impact of unprofitable customer abandonment on current customers' exit, voice, and loyalty intentions: An empirical analysis. *Journal of Services Marketing*, *26*(6), 458–470. doi: 10.1108/08876041211257936

Hajarolasvadi, N., & Demirel, H. (2019). 3D CNN-based speech emotion recognition using K-Means clustering and spectrograms. *Entropy*, *21*(5), 479. doi: 10.3390/e21050479

Han, Q., Yang, Z., Lin, H., & Qin, T. (2024). Let topic flow: A unified topic-guided segment-wise dialogue summarization framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*, 2021–2032. doi: 10.1109/TASLP.2024.3374112

Han, S., & Anderson, C. K. (2020). Customer motivation and response bias in online reviews. *Cornell Hospitality Quarterly*, *61*(2), 142–153. doi: 10.1177/1938965520902012

Hareli, S., & Rafaeli, A. (2008). Emotion cycles: On the social influence of emotion in organizations. *Research in Organizational Behavior*, *28*, 35–59. doi: 10.1016/j.riob.2008.04.007

Hathaway, B. A., Emadi, S. M., & Deshpande, V. (2021). Personalized priority policies in call centers using past customer interaction information. *Management Science*, *68*(4), 2806–2823. doi: 10.1287/mnsc.2021.4021

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics* (Vol. 2, pp. 539–545). Nantes, France: Association for Computational Linguistics. doi: 10.3115/992133.992154

Hekman, D. R., Aquino, K., Owens, B. P., & Mitchell, T. R. (2017). An examination of whether and how racial and gender biases influence customer satisfaction. *Academy of Management Journal*, *53*(2), 238–264. doi: 10.5465/amj.2010.49388763

Henkel, A. P., Bromuri, S., Iren, D., & Urovi, V. (2020). Half human, half machine - Augmenting service employees with AI for interpersonal emotion regulation. *Journal of Service Management*, *31*(2), 247–265. doi: 10.1108/JOSM-05-2019-0160

Henkel, A. P., Čaić, M., Blaurock, M., & Okan, M. (2020). Robotic transformative service research: Deploying social robots for consumer well-being during COVID-19 and beyond. *Journal of Service Management*, *31*(6), 1131–1148. doi: 10.1108/JOSM-05-2020-0145

Hillmer, S., Hillmer, B., & McRoberts, G. (2004). The real costs of turnover: Lessons from a call center. *Human Resource Planning*, *27*(3), 34–42.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. doi: 10.1126/science.aaa8685

Hofer, M., Obraczka, D., Saeedi, A., Köpcke, H., & Rahm, E. (2024). Construction of knowledge graphs: Current state and challenges. *Information*, *15*(8). doi: 10.3390/info15080509

Holman, D., Chissick, C., & Totterdell, P. (2002). The effects of performance monitoring on emotional labor and well-being in call centers. *Motivation and Emotion*, *26*, 57–81. doi: 10.1023/A:1015194108376

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, *37*(1), 81–88. doi: 10.1080/02763869.2018.1404391

Hsu, H.-H., Chen, T.-C., Chan, W.-T., & Chang, J.-K. (2016). Performance evaluation of call center agents by neural networks. In *2016 30th International Conference on Advanced Information Networking and Applications Workshop* (Vol. 2016, pp. 964–968). doi: 10.1109/WAINA.2016.126

Huang, L. L., Chen, R. P., & Chan, K. W. (2024). Pairing up with anthromorphized artificial agents: Leveraging employee creativity in service encounters. *Journal of the Academy of Marketing Science*, *52*, 955–975. doi: 10.1007/s11747-024-01017-w

Huang, M.-H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, *21*(2), 155–172. doi: 10.1177/1094670517752459

Huang, S., Peng, W., Li, J., & Lee, D. (2013). Sentiment and topic analysis on social media: A multi-task multi-label classification approach. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 172–181). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2464464.2464512

Huang, Y., & Gursoy, D. (2024). How does AI technology integration affect employees' proactive service behaviors? A transactional theory of stress perspective. *Journal of Retailing and Consumer Services*, *77*, 103700. doi: 10.1016/j.jretconser.2023.103700

Huguet Cabot, P.-L., & Navigli, R. (2021). REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2370–2381). Punta Cana, Dominican Republic: Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.204

IBM. (2024). *AI in action.* Research report. (Available at: `https://www.ibm.com/think/reports/ai-in-action`)

Iftikhar, S., & Alsamhi, S. H. (2025). Enhancing sustainability in LLM training: Leveraging federated learning and parameter-efficient fine-tuning. *IEEE Transactions on Sustainable Computing*, 1–18. doi: 10.1109/TSUSC.2025.3592043

Ilk, N., Shang, G., & Goes, P. (2020). Improving customer routing in contact centers: An automated triage design based on text analytics. *Journal of Operations Management*, *66*(5), 553–577. doi: 10.1002/joom.1084

Iren, D., Yildirim, E., & Shingjergji, K. (2023). Ethical risks, concerns, and practices of affective computing: A thematic analysis. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (pp. 1–4). doi: 10.1109/ACIIW59127.2023.10388171

Istanbulluoglu, D., & Oz, S. (2023). Service recovery via Twitter: An exploration of responses to consumer complaints. *Accounting Perspectives*, *22*(4), 435–460. doi: 10.1111/1911-3838.12339

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*, 685–695. doi: 10.1007/s12525-021-00475-2

Jarvenpaa, S. L., & Välikangas, L. (2025). Organizational learning lens: Does intelligent technology make organizations more or less intelligent? *Strategic Organization*. (In press) doi: 10.1177/14761270251350678

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., … Sayed, W. E. (2023). *Mistral 7B.* Preprint arXiv:2310.06825.

Joachimiak, M. P., Miller, M. A., Caufield, J. H., Ly, R., Harris, N. L., Tritt, A., … Bouchard, K. E. (2024). The artificial intelligence ontology: LLM-assisted construction of AI concept hierarchies. *Applied Ontology*, *19*(4), 408–418. doi: 10.1177/15705838241304103

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(27). doi: 10.1186/s40537-019-0192-5

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. doi: 10.1126/science.aaa8415

Joumblatt, D., Chandrashekar, J., Kveton, B., Taft, N., & Teixeira, R. (2013). Predicting user dissatisfaction with Internet application performance at end-hosts. In *2013 Proceedings IEEE INFOCOM* (pp. 235–239). doi: 10.1109/INFCOM.2013.6566770

Kakas, A., Mancarella, P., Sadri, F., Stathis, K., & Toni, F. (2008). Computational logic foundations of KGP agents. *Journal of Artificial Intelligence Research*, *33*, 285–348. doi: 10.1613/jair.2596

Kanchinadam, T., Meng, Z., Bockhorst, J., Singh, V., & Fung, G. (2021). *Graph neural networks to predict customer satisfaction following interactions with a corporate call center.* Retrieved from `https://arxiv.org/abs/2102.00420`

Karakus, B., & Aydin, G. (2016). Call center performance evaluation using big data analytics. In *2016 International Symposium on Networks, Computers and Communications (ISNCC)* (pp. 1–6). doi: 10.1109/ISNCC.2016.7746116

Karmakar, P., Teng, S. W., & Lu, G. (2024). Thank you for attention: A survey on attention-based artificial neural networks for automatic speech recognition. *Intelligent Systems with Applications*, *23*, 200406. doi: 10.1016/j.iswa.2024.200406

Kashima, Y., McKintyre, A., & Clifford, P. (1998). The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology*, *1*(3), 289–313. doi: 10.1111/1467-839X.00019

Katsikeas, C. S., Morgan, N. A., Leonidou, L. C., & Hult, G. T. M. (2016). Assessing performance outcomes in marketing. *Journal of Marketing*, *80*(2), 1–20. doi: 10.1509/jm.15.0287

Kazanci, N. (2025). Extended topic classification utilizing LDA and BERTopic: A call center case study on robot agents and human agents. *Applied Intelligence*, *55*(360). doi: 10.1007/s10489-024-06106-5

Keith, J. E., Lee, D.-J., & Gravois Leem, R. (2004). The effect of relational exchange between the service provider and the customer on the customer's perception of value. *Journal of Relationship Marketing*, *3*(1), 3–33. doi: 10.1300/J366v03n01_02

Kies, A., De Keyser, A., Jaramillo, S., Li, J., Tang, Y. E., & Ud Din, I. (2025). Wired for work: Brain-computer interfaces' impact on frontline employees' well-being. *Journal of Service Management*, *36*(1), 1–26. doi: 10.1108/JOSM-03-2024-0098

Kim, J. J., Kim, S. S., Ayalew, Z. A., Chua, B.-L., Han, H., & Lee, J. (2025). Do employees and customers understand their new roles for collaborative value co-creation in technology-driven service settings? *International Journal of Hospitality Management*, *128*, 104199. doi: 10.1016/j.ijhm.2025.104199

Kim, S. N., Cavedon, L., & Baldwin, T. (2010). Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 862–871).

Kim, Y., Levy, J., & Liu, Y. (2020). Speech sentiment and customer satisfaction estimation in socialbot conversations. In *Interspeech.*

Klasmeyer, G., & Sendlmeier, W. F. (2000). Voice and emotional states. In W. J. Hardcastle & J. Laver (Eds.), *Voice Quality Measurement* (pp. 339–357). London: Whurr Publishers. (Chapter 15)

Ko, Y. H., Hsu, P.-Y., Liu, Y.-C., & Yang, P.-C. (2022). Confirming customer satisfaction with tones of speech. *IEEE Access*, *10*, 83236–83248. doi: 10.1109/ACCESS.2022.3196733

Kommineni, V. K., König-Ries, B., & Samuel, S. (2024). *From human experts to machines: An LLM supported approach to ontology and knowledge graph construction.* Retrieved from `https://arxiv.org/abs/2403.08345`

Kraus, S., Oshrat, Y., Aumann, Y., Hollander, T., Maksimov, O., Ostroumov, A., & Shechtman, N. (2024). Customer service combining human operators and virtual agents: A call for multidisciplinary AI research. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(13), 15393–15401. doi: 10.1609/aaai.v37i13.26795

Krishnan, H., Elayidom, M. S., & Santhanakrishan, T. (2017). Emotion detection of Tweets using naïve Bayes classifier. *International Journal of Technology Science and Research*, *4*(11), 457–462.

Kumar, V., Chattaraman, V., Neghina, C., Skiera, B., Aksoy, L., Buoye, A., & Henseler, J. (2013). Data-driven services marketing in a connected world. *Journal of Service Management*, *24*(3), 330–352. doi: 10.1108/09564231311327021

Kusche, I. (2024). Possible harms of artificial intelligence and the EU AI act: Fundamental rights and risk. *Journal of Risk Research*, 1–14. doi: 10.1080/13669877.2024.2350720

Labat, S., Demeester, T., & Hoste, V. (2024). EmoTwiCS: A corpus for modelling emotion trajectories in Dutch customer service dialogues on Twitter. *Language Resources and Evaluation*, *58*, 505–546. doi: 10.1007/s10579-023-09700-0

Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, *103*(9), 1449–1477. doi: 10.1109/JPROC.2015.2460697

Lapré, M. A. (2011). Reducing customer dissatisfaction: How important is learning to reduce service failure? *Production and Operations Management*, *20*(4), 491–507. doi: 10.1111/j.1937-5956.2010.01149.x

Larivière, B., Bowen, D., Andreassen, T. W., Kunz, W., Sirianni, N. J., Voss, C., … De Keyser, A. (2017). "Service encounter 2.0": An investigation into the roles of technology, employees and customers. *Journal of Business Research*, *79*, 238–246. doi: 10.1016/j.jbusres.2017.03.008

Larrouy-Maestri, P., Poeppel, D., & Pell, M. D. (2024). The sound of emotional prosody: Nearly 3 decades of research and future directions. *Perspectives on Psychological Science*, *20*(4), 623–638. doi: 10.1177/17456916231217722

Laskar, M. T. R., Chen, C., Fu, X.-Y., Azizi, M., Bhushan, S., & Corston-Oliver, S. (2023). *AI coach assist: An automated approach for call recommendation in contact centers for agent coaching.* Retrieved from `https://arxiv.org/abs/2305.17619`

Le, K. B., Sajtos, L., Kunz, W. H., & Fernandez, K. V. (2024). The future of work: Understanding the effectiveness of collaboration between human and digital employees in service. *Journal of Service Research*, *28*(1), 186–205. doi: 10.1177/10946705241229419

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. doi: 10.1038/nature14539

Ledro, C., Nosella, A., & Vinelli, A. (2022). Artificial intelligence in customer relationship management: Literature review and future research directions. *Journal of Business & Industrial Marketing*, *37*(13), 48–63. doi: 10.1108/JBIM-07-2021-0332

Lee, A. V., Moriarty, J. P., Borgstrom, C., & Horwitz, L. (2010). What can we learn from patient dissatisfaction? Analysis of dissatisfying events at an academic medical center. *J Hosp Med.*, *5*(9), 514–520. doi: 10.1002/jhm.861

Lee, I.-C., Lu, J.-F. R., Fu, C.-W., & Teng, C.-I. (2017). Why can some service employees provide service of a consistently high quality while others cannot? *Service Science*, *9*(2), 167–180. doi: 10.1287/serv.2016.0171

Lee, L., & Madera, J. M. (2019). Faking it or feeling it: The emotional displays of surface and deep acting on stress and engagement. *International Journal of Contemporary Hospitality Management*, *31*(4), 1744–1762. doi: 10.1108/IJCHM-05-2018-0405

Lee, M. C., Scheepers, H., Liu, A. K., & Ngai, E. W. (2023). The implementation of artificial intelligence in organizations: A systematic literature review. *Information & Management*, *60*(5), 102225. doi: 10.1016/j.im.2023.103816

Legros, B. (2021a). Agents' self-routing for blended operations to balance inbound and outbound services. *Production and Operations Management*, *30*(10), 3599–3614. doi: 10.1111/poms.13452

Legros, B. (2021b). Routing analyses for call centers with human and automated services. *International Journal of Production Economics*, *240*, 108247. doi: 10.1016/j.ijpe.2021.108247

Leiño Calleja, D., Schepers, J., & Nijssen, E. J. (2025). Hybrid human-robot teams in the frontline: Automated social presence and the role of corrective interrogation. *Journal of Service Management*, *ahead-of-print*(ahead-of-print). doi: 10.1108/JOSM-11-2023-0470

Lemmens, A., & Gupta, S. (2020). Managing churn to maximize profits. *Marketing Science*, *39*(5), 956–973. doi: 10.1287/mksc.2020.1229

Lervik Olsen, L., Witell, L., & Gustafsson, A. (2014). Turning customer satisfaction measurements into action. *Journal of Service Management*, *25*(4), 556–571. doi: 10.1108/JOSM-01-2014-0025

Levi-Bliech, M., Pliskin, N., & Fink, L. (2020). Implementing a sales support app to complement face-to-face interaction: An empirical investigation of business value. *Journal of Organizational Computing and Electronic Commerce*, *30*(3), 266–278. doi: 10.1080/10919392.2020.1750932

Li, J., Sun, A., Han, J., & Li, C. (2022). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *34*(1), 50–70. doi: 10.1109/TKDE.2020.2981314

Li, W., Shao, W., Ji, S., & Cambria, E. (2022). BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing*, *467*, 73–82. doi: 10.1016/j.neucom.2021.09.057

Li, Y., Xing, A., & Terui, N. (2023). Modeling customer satisfaction's impact on loyalty: Insights for customer-centric resource allocation. *Service Science*, *15*(2), 107–128. doi: 10.1287/serv.2022.0313

Libai, B., Bart, Y., Gensler, S., Hofacker, C. F., Kaplan, A., Kötterheinrich, K., & Kroll, E. B. (2022). Brave new world? On AI and the management of customer relationships. *Journal of Interactive Marketing*, *51*(1), 44–56. doi: 10.1016/j.intmar.2020.04.002

Lin, H., Zhu, J., Xiang, L., Zhai, F., Zhou, Y., & Zhang, J. (2023). Topic-oriented dialogue summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 1791–1810. doi: 10.1109/TASLP.2023.3271118

Lin, H.-F. (2025). An integrated model examining frontline employee willingness to work with retail service robots. *Journal of Service Theory and Practice*, *35*(3), 464–482. doi: 10.1108/JSTP-07-2024-0231

Lin, X., Wang, X., Shao, B., & Taylor, J. (2024). How chatbots augment human intelligence in customer services: A mixed-methods study. *Journal of Management Information Systems*, *41*(4), 1016–1041. doi: 10.1080/07421222.2024.2415773

Liu, X., Sun, J., Lei, A., & Zhu, J. (2024). Research and applications of large language models for converting unstructured data into structured data. In *2024 3rd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)* (pp. 305–308). doi: 10.1109/CBASE64041.2024.10824634

Liu, Y., Chi, M., & Sun, Q. (2024). Sarcasm detection in hotel reviews: A multimodal deep learning approach. *Journal of Hospitality and Tourism Technology*, *15*(4), 519–533. doi: 10.1108/JHTT-04-2023-0098

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach.* Retrieved from https://arxiv.org/abs/1907.11692

Liu, Z., Long, C., Lu, X., Hu, Z., Zhang, J., & Wang, Y. (2019). Which channel to ask my question?: Personalized customer service request stream routing using deep reinforcement learning. *IEEE Access*, *7*, 107744–107756. doi: 10.1109/ACCESS.2019.2932047

Lo, A., Jiang, A. Q., Li, W., & Jamnik, M. (2024). End-to-end ontology learning with large language models. In *Advances in Neural Information Processing Systems* (Vol. 37, pp. 87184–87225). doi: 10.17863/CAM.116988

Lu, V. N., Wirtz, J., Kunz, W. H., Paluch, S., Gruber, T., Martins, A., & Patterson, G., Paul. (2020). Service robots, customers and service employees: What can we learn from the academic literature and where are the gaps? *Journal of Service Theory and Practice*, *30*(3), 361–391. doi: 10.1108/JSTP-04-2019-0088

Lukitasari, S. D., & Hidayat, F. (2020). Deep learning-based complaint classification for Indonesia telecommunication company's call center. In *Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar, MSCEIS 2019, 12 October 2019, Bandung, West Java, Indonesia.* EAI. doi: 10.4108/eai.12-10-2019.2296518

Luong, M.-T., Pham, H., & Manning, C. D. (2015). *Effective approaches to attention-based neural machine translation.* Retrieved from https://arxiv.org/abs/1508.04025 doi: 10.48550/arXiv.1508.04025

Luque, J., Segura, C., Sánchez, A., Umbert, M., & Galindo, L. A. (2017). The role of linguistic and prosodic cues on the prediction of self-reported satisfaction in contact centre phone calls. In *Interspeech* (pp. 2346–2350). doi: http://dx.doi.org/10.21437/Interspeech.2017-424

Lyu, C., Wu, M., Wang, L., Huang, X., Liu, B., Du, Z., … Tu, Z. (2023). *Macaw-LLM: Multi-modal language modeling with image, audio, video, and text integration.* Retrieved from https://arxiv.org/abs/2306.09093

Ma, C., & Ye, J. (2022). Linking artificial intelligence to service sabotage. *The Service Industries Journal*, *42*(13-14), 1054–1074. doi: 10.1080/02642069.2022.2092615

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing - Connecting computing power to human insights. *International Journal of Research in Marketing*, *37*(3), 481–504. doi: 10.1016/j.ijresmar.2020.04.005

Ma, X., Deng, C., Du, D., & Pei, Q. (2023). An enhanced method for dialect transcription via error-correcting thesaurus. *IET Communications*, *17*(17), 1984–1997. doi: 10.1049/cmu2.12671

Maedche, A., & Staab, S. (2000). Discovering conceptual relations from text. In *Proceedings of the 14th European Conference on Artificial Intelligence* (pp. 321–325). doi: 10.5555/3006433.3006501

Maglio, P. P., & Spohrer, J. (2008). Fundamentals of service science. *Journal of the Academy of Marketing Science*, *36*, 18–20. doi: 10.1007/s11747-007-0058-9

Manderscheid, E., & Lee, M. (2023). Predicting customer satisfaction with soft labels for ordinal classification. In S. Sitaram, B. Beigman Klebanov, & J. D. Williams (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Vol. 5, pp. 652–659). Toronto, Canada: Association for Computational Linguistics.

Manno, A., Rossi, F., Smriglio, S., & Cerone, L. (2023). Comparing deep and shallow neural networks in forecasting call center arrivals. *Soft Computing*, *27*, 12943–12957. doi: 10.1007/s00500-022-07055-2

Marinova, D., de Ruyter, K., Huang, M.-H., Meuter, M. L., & Challagalla, G. (2016). Getting smart: Learning from technology-empowered frontline interactions. *Journal of Service Research*, *20*(1), 29–42. doi: 10.1177/1094670516679273

Marr, B. (2024, January 26). *How generative AI is revolutionizing customer service.* Forbes. (Accessed: 2024-10-16)

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., … Flach, P. (2021). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. doi: 10.1109/TKDE.2019.2962680

Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2024). Prompt engineering in large language models. In I. J. Jacob, S. Piramuthu, & P. Falkowski-Gilski (Eds.), *Data Intelligence and Cognitive Informatics* (pp. 387–402). Singapore: Springer Nature Singapore.

Marín Díaz, G., Gómez Medina, R., & Aijón Jiménez, J. A. (2025). A methodological framework for business decisions with explainable AI and the analytic hierarchical process. *Processes*, *13*(1), 102–126. doi: 10.3390/pr13010102

Mattila, A. S., & Enz, C. A. (2002). The role of emotions in service encounters. *Journal of Service Research*, *4*(4), 268–277. doi: 10.1177/1094670502004004004

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). *A Proposal for the Dartmouth summer research project on artificial intelligence.* Dartmouth College. Retrieved from https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

McDonald, R., & Nivre, J. (2011). Analyzing and integrating dependency parsers. *Computational Linguistics*, *37*(1), 197–230. doi: 10.1162/coli_a_00039

McKinsey & Company. (2024). *AI mastery in customer care: Raising the bar for quality assurance.* https://www.mckinsey.com/capabilities/operations/our-insights/operations-blog/ai-mastery-in-customer-care-raising-the-bar-for-quality-assurance. (Accessed: 2025-08-06)

McLean, G., & Osei-Frimpong, K. (2017). Examining satisfaction with the experience during a live chat service encounter-implications for website providers. *Computers in Human Behavior*, *76*, 494–508. doi: 10.1016/j.chb.2017.08.005

Meinzer, S., Jensen, U., Thamm, A., Hornegger, J., & Eskofier, B. M. (2016). Can machine learning techniques predict customer dissatisfaction? A feasibility study for the automotive industry. *Artificial Intelligence Research*, *6*(1), 80–90. doi: 10.5430/air.v6n1p80

Mele, A. R. (1989). Intention, belief, and intentional action. *American Philosophical Quarterly*, *26*(1), 19–30.

Microsoft. (2024). *Microsoft Learn: Azure OpenAI Service Models.* https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models. (Accessed: 2024-07-18)

Microsoft. (2024). *Phi open models.* https://azure.microsoft.com/en-us/products/phi. (Accessed: September 29th, 2024)

Mihindukulasooriya, N., Tiwari, S., Enguix, C. F., & Lata, K. (2023). Text2KGBench: A benchmark for ontology-driven knowledge graph generation from text. In *The Semantic Web – ISWC 2023* (pp. 247–265). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-47243-5_14

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. doi: 10.1037/h0043158

Mittal, V., & Frennea, C. (2010). *Customer satisfaction: A strategic review and guidelines for managers* (Research Report). Cambridge, MA: Marketing Science Institute. Retrieved from https://ssrn.com/abstract=2345469

Mohammed, R. A. (2017). Using personalized model to predict traffic jam in inbound call center. *EAI Endorsed Transactions*, *4*(12), 1–5. doi: 10.4108/eai.18-1-2017.152101

Moliner-Tena, M. A., Callarisa-Fiol, L. J., Sánchez-García, J., & Rodríguez-Artola, R. M. (2024). Co-creation 5.0: The frontline employee-robot team and firms' outcomes. The Tin Woodman paradox. *Journal of Innovation & Knowledge*, *9*(3), 100534. doi: 10.1016/j.jik.2024.100534

Montgomery, L., Damian, D., Bulmer, T., & Quader, S. (2018). Customer support ticket escalation prediction using feature engineering. *Requirements Engineering*, *23*, 333-355. doi: 10.1007/s00766-018-0292-3

Montobbio, F., Staccioli, J., Virgillito, M. E., & Vivarelli, M. (2023). The empirics of technology, employment and occupations: Lessons learned and challenges ahead. *Journal of Economic Surveys*, *38*(5), 1622–1655. doi: 10.1111/joes.12601

Moshavi, D. (2006). He said, she said: Gender bias and customer satisfaction with phone-based service encounters. *Journal of Applied Social Psychology*, *34*(1), 162–176. doi: 10.1111/j.1559-1816.2004.tb02542.x

Mukherjee, A., Burnham, T., & King, D. (2021). Anticipated firm interaction can bias expressed customer satisfaction. *Journal of Retailing and Consumer Services*, *59*, 102379. doi: 10.1016/j.jretconser.2020.102379

Mustak, M., Salminen, J., Plé, L., & Wirtz, J. (2021). Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *Journal of Business Research*, *124*, 389–404. doi: 10.1016/j.jbusres.2020.10.044

Mustaqeem, & Kwon, S. (2021). MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Systems With Applications*, *167*, 114177. doi: 10.1016/j.eswa.2020.114177

Mustaqeem, Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access*, *8*, 79861–79875. doi: 10.1109/ACCESS.2020.2990405

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A review of evaluation metrics in machine learning algorithms. In R. Silhavy & P. Silhavy (Eds.), *Artificial Intelligence Application in Networks and Systems* (pp. 15–25). Cham: Springer International Publishing. doi: 10.1007/978-3-031-35314-7_2

Namli, O. H., Yanik, S., Nouri, F., Serap Şengör, N., Koyuncu, Y. M., & Uçar, O. B. (2021). A neural networks approach to predict call center calls of an internet service provider. *Journal of Intelligent & Fuzzy Systems*, *42*(1), 503–515. doi: 10.3233/JIFS-219207

Navarro, J. (2023). *Consequences of bad customer experience in the United States and selected countries in Western Europe as of September 2021.* `https://www.statista.com/statistics/1358520/bad-customer-experience-consequences/`. Statista Research Department. (Accessed: 2024-10-09)

Navigli, R., Velardi, P., & Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, *18*(1), 22–31. doi: 10.1109/MIS.2003.1179190

Nishant, R., Schneckenberg, D., & Ravishankar, M. (2023). The formal rationality of artificial intelligence-based algorithms and the problem of bias. *Journal of Information Technology*, *39*(1), 19–40. doi: 10.1177/02683962231176842

Noroozi, F., Corneanu, C. A., Kaminska, D., Sapinski, T., Escalera, S., & Anbarjafari, G. (2021). Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, *12*(2), 505–523. doi: 10.1109/TAFFC.2018.2874986

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., … Staab, S. (2020). Bias in data-driven artificial intelligence systems - An introductory survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), e1356. doi: 10.1002/widm.1356

Nyberg, D. (2009). Computers, customer service operatives and cyborgs: Intra-actions in call centers. *Organization Studies*, *30*(1), 1181–1199. doi: 10.1177/0170840609337955

Obinna Iheme, L., & Ozan, c. (2022). A novel semi-supervised framework for call center agent malpractice detection via neural feature learning. *Expert Systems with Applications*, *208*, 118173. doi: 10.1016/j.eswa.2022.118173

Odekerken-Schröder, G., Mennens, K., Steins, M., & Mahr, D. (2022). The service triad: An empirical study of service robots, customers and frontline employees. *Journal of Service Management*, *33*(2), 246–292. doi: 10.1108/JOSM-10-2020-0372

Oder, N., & Béland, D. (2025). Artificial intelligence, emotional labor, and the quest for sociological and political imagination among low-skilled workers. *Policy and Society*, *44*(1), 116–128. doi: 10.1093/polsoc/puae034

Oliver, R. L. (1996). *Satisfaction: A behavioral perspective on the consumer*. New York: McGraw-Hill.

O'Neill, A. (2025). *Share of economic sectors in the global gross domestic product from 2014 to 2024.* https://www.statista.com/statistics/256563/share-of-economic-sectors-in-the-global-gross-domestic-product/. Statista Research Department. (Accessed: 2025-07-30)

OpenAI. (2023). *ChatGPT can now see, hear, and speak.* OpenAI Blog. Retrieved from https://openai.com/blog/chatgpt-can-now-see-hear-and-speak (Accessed: September 25, 2023)

Oraby, S., Bhuiyan, M., Gundecha, P., Mahmud, J., & Akkiraju, R. (2019). Modeling and computational characterization of Twitter customer service conversations. *ACM Transactions on Interactive Intelligent Systems*, *9*(2-3). doi: 10.1145/3213014

Oraby, S., Gundecha, P., Mahmud, J., Bhuiyan, M., & Akkiraju, R. (2017). "How may I help you?": Modeling Twitter customer service conversations using fine-grained dialogue acts. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (pp. 343–355). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3025171.3025191

Osman, M. A., Mohd Noah, S. A., & Saad, S. (2022). Ontology-based knowledge management tools for knowledge sharing in organization- A review. *IEEE Access*, *10*, 43267–43283. doi: 10.1109/ACCESS.2022.3163758

Ostrom, A. L., Parasuraman, A., Bowen, D. E., Patrício, L., & Voss, C. A. (2015). Service research priorities in a rapidly changing context. *Journal of Service Research*, *18*(2), 127–159. doi: 10.1177/1094670515576315

Oztaysi, B., Onar, S. C., Kahraman, C., & Gok, M. (2020). Call center performance measurement using intuitionistic fuzzy sets. *Journal of Enterprise Information Management*, *33*(6), 1647–1668. doi: 10.1108/JEIM-04-2017-0050

Pacella, M., Vasco, P., Papadia, G., & Giliberti, V. (2024). An assessment of digitalization techniques in contact centers and their impact on agent performance and well-being. *Sustainability*, *16*(2), 714. doi: 10.3390/su16020714

Pachet, F., & Roy, P. (2009). Analytical features: A knowledge-based approach to audio feature generation. *EURASIP Journal on Audio, Speech, and Music Processing*(153017). doi: 10.1155/2009/153017

Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press.

Papadia, G., Pacella, M., & Giliberti, V. (2022). Topic modeling for automatic analysis of natural language: A case study in an Italian customer support center. *Algorithms*, *15*(6), 204. doi: 10.3390/a15060204

Papadia, G., Pacella, M., Perrone, M., & Giliberti, V. (2023). A comparison of different topic modeling methods through a real case study of Italian customer care. *Algorithms*, *16*(2), 94. doi: 10.3390/a16020094

Paprzycki, M., Abraham, A., Guo, R., & Mukkamala, S. (2004). Data mining approach for analyzing call center performance. *Innovations in Applied Artificial Intelligence*, *3029*, 1092–1101. doi: 10.1007/978-3-540-24677-0_112

Park, E., Lee, D., Han, Y., Diefendorff, J., & Lee, U. (2024). Hide-and-seek: Detecting workers' emotional workload in emotional labor contexts using multimodal sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(3), 1–28. doi: 10.1145/367859

Park, K., Cha, M., & Rhim, E. (2018). Positivity bias in customer satisfaction ratings. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18* (pp. 631–638). ACM Press. doi: 10.1145/3184558.3186579

Park, T.-Y., & Shaw, J. D. (2013). Turnover rates and organizational performance: A meta-analysis. *Journal of Applied Psychology*, *98*(2), 268–309. doi: 10.1037/a0030723

Park, Y. (2011). Automatic call quality monitoring using cost-sensitive classification. In *Interspeech* (pp. 3085–3088).

Park, Y., & Gates, S. C. (2009). Towards real-time measurement of customer satisfaction using automatically generated call transcripts. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (p. 1387-1396). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/1645953.1646128

Parra-Gallego, L. F., & Orozco-Arroyave, J. R. (2022). Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments. *Digital Signal Processing*, *120*, 103286. doi: 10.1016/j.dsp.2021.103286

Payne, A., & Frow, P. (2005). A strategic framework for customer relationship management. *Journal of Marketing*, *69*(4), 167–176. doi: 10.1509/jmkg.2005.69.4.167

Phillips, C., Odekerken-Schröder, G., Russell-Bennett, R., Steins, M., Mahr, D., & Letheren, K. (2025). Service robot-employee task allocation strategies: Well-being within the intrusion challenge. *Journal of Service Management*, *ahead-of-print*(ahead-of-print). doi: 10.1108/JOSM-11-2023-0466

Picard, R. W. (1999). Affective computing for HCI. In *Proceedings of the HCI Conference* (pp. 829–833).

Pine, B. J., Victor, B., & Boynton, A. C. (1993). Making mass customization work. *Harvard Business Review*, *71*(5), 108–111.

Plaza, M., Pawlik, L., & Deniziak, S. (2021). Call transcription methodology for contact center systems. *IEEE Access*, *9*, 110975–110988. doi: 10.1109/ACCESS.2021.3102502

Plutchik, R., & Kellerman, H. (Eds.). (1980). *Emotion: Theory, research, and experience. Vol. 1: Theories of Emotion*. New York: Academic Press.

Poczeta, K., Plaza, M., Zawadzki, M., Michno, T., & Krechowicz, M. (2024). Analysis of the retaining strategies for multi-label text message classification in call/contact center systems. *Scientific Reports*, *14*(10093). doi: 10.1038/s41598-024-60697-0

Pontes, M. C., & O'Brien Kelly, C. (2000). The identification of inbound call center agents' competencies that are related to callers' repurchase intentions. *Journal of Interactive Marketing*, *14*(3), 41–49. doi: 10.1002/1520-6653(200022)14:3<41::AID-DIR3>3.0.CO;2-M

Ponzetto, S. P., & Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence* (Vol. 2, pp. 1440–1445). doi: 10.5555/1619797.1619876

Poots, J., Morgan, J., Woolf, J., & Curcuruto, M. (2024). Identifying system adaptations to overcome technology-based workflow challenges in a telephone triage organization. *Applied Ergonomics*, *121*, 104365. doi: 10.1016/j.apergo.2024.104365

Posselt, T., & Gerstner, E. (2005). Pre-sale vs. post-sale e-satisfaction: Impact on repurchase intention and overall satisfaction. *Journal of Interactive Marketing*, *19*(4), 35–47. doi: 10 .1002/dir.20048

Prentice, C., Lopes, S. D., & Wang, X. (2019). Emotional intelligence or artificial intelligence - An employee perspective. *Journal of Hospitality Marketing & Management*, *29*(4), 377–403. doi: 10.1080/19368623.2019.1647124

Presbitero, A. (2016). It's not all about language ability: Motivational cultural intelligence matters in call center performance. *The International Journal of Human Resource Management*, *28*(11), 1547–1562. doi: 10.1080/09585192.2015.1128464

Pérez-Toro, P. A., Vásquez-Correa, J. C., Bocklet, T., Nöth, E., & Orozco-Arroyave, J. R. (2023). User state modeling based on the arousal-valence plane: Applications in customer satisfaction and health-care. *IEEE Transactions on Affective Computing*, *14*(2), 1533–1546. doi: 10.1109/TAFFC.2021.3112543

Pöyry, E., Holopainen, J., Parvinen, P., Mattila, O., & Tuunanen, T. (2024). Design principles for virtual reality applications used in collaborative service encounters. *Journal of Service Research*, *0*(0). doi: 10.1177/10946705241266971

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning, pages = 28492–28518* (Vol. 202). PMLR.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training.* OpenAI. Retrieved from `https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf`

Ragheb, W., Azé, J., Bringay, S., & Servajean, M. (2019). *Attention-based modeling for emotion detection and classification in textual conversations.* Retrieved from `https://arxiv.org/abs/1906.07020`

Rajaobelina, L., Brun, I., & Ricard, L. (2019). A classification of live chat service users in the banking industry. *International Journal of Bank Marketing*, *37*(3), 838–857. doi: 10.1108/IJBM-03-2018-0051

Rajwadi, M., Glackin, C., Wall, J., Chollet, G., & Cannings, N. (2019). Explaining sentiment classification. In *Interspeech.* doi: 10.21437/Interspeech.2019-2743

Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning* (pp. 1–18). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rao, A. S., & Georgeff, M. P. (1995). BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multiagent Systems* (pp. 312–319).

Reeck, C., & Onuklu, N. N. Y. (2022). Interpersonal emotion regulation: Consequences for brands in customer service interactions. *Frontiers in Psychology*, *13*(872670). doi: 10.3389/fpsyg.2022.872670

Rees, D., Laramee, R. S., Brookes, P., D'Cruze, T., Smith, G. A., & Miah, A. (2021). AgentVis: Visual analysis of agent behavior with hierarchical glyphs. *IEEE Transactions of Visualization and Computer Graphics*, *27*(9), 3626–3643. doi: 10.1109/TVCG.2020.2985923

Ribeiro, H., Barbosa, B., Moreira, A. C., & Rodrigues, R. G. (2024). Determinants of churn in telecommunication services: A systematic literature review. *Management Review Quarterly*, *74*, 1327–1364. doi: 10.1007/s11301-023-00335-7

Risselada, H., Verhoef, P. C., & Bijmolt, T. H. (2010). Staying power of churn prediction models. *Journal of Interactive Marketing*, *24*(3), 198–208. doi: 10.1016/j.intmar.2010.04.002

Rivera, M., Qiu, L., Kumar, S., & Petrucci, T. (2021). Are traditional performance reviews outdated? An empirical analysis on continuous, real-time feedback in the workplace. *Information Systems Research*, *32*(2), 517–540. doi: 10.1287/isre.2020.0979

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach, 4th edition*. Pearson.

Rust, R. T., & Huang, M.-H. (2014). The service revolution and the transformation of marketing science. *Marketing Science*, *33*(2), 206–221. doi: 10.1287/mksc.2013.0836

Saberi, M., Theobald, M., Hussain, O. K., Chang, E., & Hussain, F. K. (2018). Interactive feature selection for efficient customer recognition in contact centers: Dealing with common names. *Expert Systems with Applications*, *113*, 356–376. doi: 10.1016/j.eswa.2018.07.012

Saeedizade, M. J., & Blomqvist, E. (2024). Navigating ontology development with large language models. In A. Meroño Peñuela et al. (Eds.), *The Semantic Web* (pp. 143–161). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-60626-7_8

Sailunaz, K., Dhaliwal, M., Rokne, J., & Alhajj, R. (2018). Emotion detection from text and speech: A survey. *Social Network Analysis and Mining*, *8*(28). doi: 10.1007/s13278-018-0505-2

Sandra, L., Prabowo, H., Gaol, F. L., & Isa, S. M. (2024). PersoNet: A novel framework for personality classification-based apt customer service agent selection. *IEEE Access*, *12*, 25200–25214. doi: 10.1109/ACCESS.2024.3364352

Saon, G., Ramabhadran, B., & Zweig, G. (2006). On the effect of word error rate on automated quality monitoring. In *2006 IEEE Spoken Language Technology Workshop* (pp. 106–109). doi: 10.1109/SLT.2006.326828

Schecter, A., Wowak, K. D., Berente, N., Ye, H., & Mukherjee, U. (2021). A behavioral perspective on service center routing: The role of inertia. *Journal of Operations Management*, *67*(8), 964–988. doi: 10.1002/joom.1156

Schoenmueller, V., Netzer, O., & Stahl, F. (2020). The polarity of online reviews: Prevalence, drivers and implications. *Journal of Marketing Research*, *57*(5), 853–877. doi: 10.1177/0022243720941832

Seddon, J., & Srinivasan, R. (2014). Information and ontologies: Challenges in scaling knowledge for development. *Journal of the Association for Information Science and Technology*, *65*(6), 1124–1133. doi: 10.1002/asi.23000

Segura, C., Balcells, D., Umbert, M., Arias, J., & Luque, J. (2016). Automatic speech feature learning for continuous prediction of customer satisfaction in contact center phone calls. In A. Abad et al. (Eds.), *Advances in Speech and Language Technologies for Iberian Languages* (pp. 255–265). Cham: Springer International Publishing.

Seng, K. P., & Ang, L.-M. (2018). Video analytics for customer emotion and satisfaction at contact centers. *IEEE Transactions on Human-Machine Systems*, *48*(3), 266–278. doi: 10.1109/THMS.2017.2695613

Shabanpour, A., Hou, Z., Husnoo, A., Nguyen, K. L., Yearwood, J., & Zaidi, N. (2023). Aspect-based automated evaluation of dialogues. *Knowledge-Based Systems*, *279*, 110901. doi: 10.1016/j.knosys.2023.110901

Shah, S., Ghomeshi, H., Vakaj, E., Cooper, E., & Fouad, S. (2023). A review of natural language processing in contact centre automation. *Pattern Analysis and Applications*, *26*, 823–846. doi: 10.1007/s10044-023-01182-8

Shahin, M., Chen, F. F., Hosseinzadeh, A., Maghanaki, M., & Eghbalian, A. (2024). A novel approach to voice of customer extraction using GPT-3.5 Turbo: Linking advanced NLP and lean six sigma 4.0. *The International Journal of Advanced Manufacturing technology*, *131*, 3615–3630. doi: 10.1007/s00170-024-13167-w

Sheng, M. L., Natalia, N., & Rusfian, E. Z. (2024). AI chatbot, human, and in-between: Examining the broader spectrum of technology-human interactions in driving customer-brand relationships across experience and credence services. *Psychology & Marketing*, *42*(4), 1051–1071. doi: 10.1002/mar.22165

Sheth, J. N., Jain, V., & Ambika, A. (2023). The growing importance of customer-centric support services for improving customer experience. *Journal of Business Research*, *164*, 113943. doi: 10.1016/j.jbusres.2023.113943

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., … Yang, X. (2018). A review of emotion recognition using physiological signals. *Sensors*, *18*(7), 2074. doi: 10.3390/s18072074

Sieben, I., De Grip, A., Longen, J., & Sørensen, O. (2009). Technology, selection, and training in call centers. *ILR Review*, *62*(4), 553–572. doi: 10.1177/001979390906200405

Sikveland, R. O., & Zeitlyn, D. (2017). Using prosodic cues to identify dialogue acts: Methodological challenge. *Text & Talk*, *37*(3), 1–40. doi: 10.1515/text-2017-0007

Silva, T. F. L. D., Aduna, G. F., Benamara, F., Mari, A., Li, Z., Yue, L., & Su, J. (2025). CDB: A unified framework for hope speech detection through counterfactual, desire and belief. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 4448–4463).

Albuquerque, New Mexico: Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.252

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–489. doi: 10.1038/nature16961

Souca, M. L. (2014). Customer dissatisfaction and delight: Completely different concepts, or part of a satisfaction continuum? *Management & Marketing*, *9*(1), 75–90.

Sprigg, C. A., Stride, C. B., Wall, T. D., Holman, D. J., & Smith, P. R. (2007). Work characteristics, musculoskeletal disorders, and the mediating role of psychological strain: A study of call center employees. *Journal of Applied Psychology*, *92*(5), 1456–1466. doi: 10.1037/0021-9010.92.5.1456

Statista Research Department. (n.d.). *Business services.* Statista. Retrieved from `https://www.statista.com/markets/406/topic/430/business-services/overview` (Accessed: 16 January 2025)

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Stephens, N., & Gwinner, K. P. (1998). Why don't some people complain? A cognitive-emotive process model of consumer complaint behavior. *Journal of the Academy of Marketing Science*, *26*(3), 172–189. doi: 10.1177/0092070398263001

Sun, J., Han, P., Cheng, Z., Wu, E., & Wang, W. (2020). Transformer based multi-grained attention network for aspect-based sentiment analysis. *IEEE Access*, *8*, 211152–211163. doi: 10.1109/ACCESS.2020.3039470

Sun, X. (2019). Multi-attribute intelligent queueing method for onboard call centers. *Cluster Computing*, *22*, 5207–5215. doi: 10.1007/s10586-017-1173-0

Sun, X., & Liu, W. (2023). Expanding service capabilities through an on-demand workforce. *Operations Research*, *73*(1), 363–384. doi: 10.1287/opre.2021.0651

Swanson, S. R., & Kelley, S. W. (2001). Service recovery attributions and word-of-mouth intentions. *European Journal of Marketing*, *35*(1/2), 194–211. doi: 10.1108/03090560110363463

Sørum, H., & Presthus, W. (2020). Dude, where's my data? The GDPR in practice, from a consumer's point of view. *Information Technology & People*, *34*(3), 912–929. doi: 10.1108/ITP-08-2019-0433

Tahir, A. H., Adnan, M., & Saeed, Z. (2024). The impact of brand image on customer satisfaction and brand loyalty: A systematic literature review. *Heliyon*, *10*(16). doi: 10.1016/j.heliyon.2024.e36254

Tamura, A., Ishikawa, K., Saikou, M., & Tsuchida, M. (2011). Extractive summarization method for contact center dialogues based on call logs. In *Proceedings of the 5th International Joint Conference on Natural Language Processing* (pp. 500–508).

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., … Zhang, C. (2024). *SALMONN: Towards generic hearing abilities for large language models.* Retrieved from `https://arxiv.org/abs/2310.13289`

Terui, N., Hasegawa, S., Chun, T., & Ogawa, K. (2011). Hierarchical Bayes modeling of the customer satisfaction index. *Service Science*, *3*(2), 127–140. doi: 10.1287/serv.3.2.127

Teubner, T., Flath, C. M., Weinhardt, C., van der Aalst, W., & Hinz, O. (2023). Welcome to the era of ChatGPT et al.: The prospects of large language models. *Business & Information Systems Engineering*, *65*, 95–101. doi: 10.1007/s12599-023-00795-x

Tombs, A. G., Russell-Bennett, R., & Ashkanasy, N. M. (2014). Recognising emotional expressions of complaining customers: A cross-cultural study. *European Journal of Marketing*, *48*(7/8), 1354–1374. doi: 10.1108/EJM-02-2011-0090

Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, *42*(9), 1600–1631. doi: 10.1002/smj.3322

Tovar, J. (2021). Call center agents' skills. *Sociolinguistic Studies*, *14*(4), 437–458. doi: 10.1558/sols.39555

Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the Longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. *Human Relations*, *4*(1), 3–38. doi: 10.1177/001872675100400101

Tudorache, T. (2019). Ontology engineering: Current state, challenges, and future directions. *Semantic Web*, *11*(1), 125–138. doi: 10.3233/SW-190382

Upamannyu, N. K., & Sankpal, S. (2014). Effect of brand image on customer satisfaction & loyalty intention and the role of customer satisfaction between brand image and loyalty intention. *Journal of Social Science Research*, *3*(2), 274–285.

Valizada, A., Akhundova, N., & Rustamov, S. (2021). Development of speech recognition systems in emergency call centers. *Symmetry*, *13*(4), 634. doi: 10.3390/sym13040634

Valle, M. A., Varas, S., & Ruz, G. A. (2012). Job performance prediction in a call center using a naive Bayes classifier. *Expert Systems with Applications*, *39*(11), 9939–9945. doi: 10.1016/j.eswa.2011.11.126

van Dolen, W., Lemmink, J., de Ruyter, K., & de Jong, A. (2002). Customer-sales employee encounters: A dyadic perspective. *Journal of Retailing*, *78*(4), 265–279. doi: 10.1016/S0022-4359(02)00067-2

van Doorn, J., Mende, M., Noble, S. M., Hulland, J., Ostrom, A. L., Grewal, D., & Petersen, J. A. (2016). Domo arigato Mr. Roboto: Emergence of automated social presence in organizational frontlines and customers' service experiences. *Journal of Service Research*, *20*(1), 43–58. doi: 10.1177/1094670516679272

Van Herck, R., Decock, S., & De Clerck, B. (2020). "Can you send us a PM please?" Service recovery interactions on social media from the perspective of organizational legitimacy. *Discourse, Context & Media*, *38*, 100445. doi: 10.1016/j.dcm.2020.100445

Van Herck, R., Decock, S., & Fastrich, B. (2022). A unique blend of interpersonal and transactional strategies in English email responses to customer complaints in a B2C setting: A move analysis. *English for Specific Purposes*, *65*, 30–48. doi: 10.1016/j.esp.2021.08.001

Van Mulken, M. (2024). What verbal de-escalation techniques are used in complaint handling? *Journal of Pragmatics*, *220*, 116–131. doi: 10.1016/j.pragma.2023.12.008

Vargo, S. L., & Lusch, R. F. (2008). Service-dominant logic: Continuing the evolution. *Journal of the Academy of Marketing Science*, *36*, 1–10. doi: 10.1007/s11747-007-0069-6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)* (Vol. 30, pp. 5998–6008). Curran Associates Inc.

Verduyn, P., & Lavrijsen, S. (2015). Which emotions last longest and why: The role of event importance and rumination. *Motivation and Emotion*, *39*, 119–127. doi: 10.1007/s11031-014-9445-y

Veres, S. M., & Luo, J. (2004). A class of BDI agent architectures for autonomous control. In *2004 43rd IEEE Conference on Decision and Control* (Vol. 5, pp. 4746–4751). doi: 10.1109/CDC.2004.1429540

Verma, J. P., Agrawal, S., Patel, B., & Patel, A. (2016). Big data analytics: Challenges and applications for text, audio, video, and social media data. *International Journal on Soft Computing*, *5*(1).

Vo, N. N., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. *Knowledge-Based Systems*, *212*, 106586. doi: 10.1016/j.knosys.2020.106586

Völker, J., Hitzler, P., & Cimiano, P. (2007). Acquisition of OWL DL axioms from lexical resources. In *The Semantic Web: Research and Applications* (Vol. 4519, pp. 670–685). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-540-72667-8_47

Waelbers, B., Bromuri, S., & Henkel, A. P. (2022). Comparing neural networks for speech emotion recognition in customer service interactions. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). doi: 10.1109/IJCNN55064.2022.9892165

Walczuch, R., Lemmink, J., & Streukens, S. (2007). The effect of service employees' technology readiness on technology acceptance. *Information & Management*, *44*, 206–215. doi: 10.1016/j.im.2006.12.005

Walker, D. D., Kim, S. K. I., van Jaarsveld, D. D., Restubog, S. L. D., Marrone, M., Lagios, C., & Mehdipour, A. M. (2023). It takes two to tango: A multidisciplinary bibliometric review across six decades of dyadic service encounter research. *Journal of Service Management*, *34*(5), 970–994. doi: 10.1108/JOSM-08-2022-0286

Walker, M. A., Passonneau, R., & Boland, J. E. (2001). Quantitative and qualitative evaluation of DARPA communicator spoken dialogue systems. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (pp. 515–522).

Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J., & Tarokh, V. (2020). Speech emotion recognition with dual-sequence LSTM architecture. In *Proceedings of ICASSP '20'* (pp. 6474–6478). doi: 10.1109/ICASSP40776.2020.9054629

Wasserman, P. D., & Schwartz, T. (1988). Neural networks. II. What are they and why is everybody so interested in them now? *IEEE Expert*, *3*(1), 10–15. doi: 10.1109/64.2091

Wei, Y., Lu, W., Cheng, Q., Jiang, T., & Liu, S. (2022). How humans obtain information from AI: Categorizing user messages in human-AI collaborative conversations. *Information Processing & Management*, *59*(2), 102838. doi: 10.1016/j.ipm.2021.102838

White, C., & Roos, V. (2005). Core competencies of a call centre agent. *SA Journal of Human Resource Management*, *3*(2), 41–47. doi: 10.4102/sajhrm.v3i2.63

Wirtz, J., Kunz, W. H., Hartley, N., & Tarbit, J. (2022). Corporate digital responsibility in service firms and their ecosystems. *Journal of Service Research*, *26*(2), 173–190. doi: 10.1177/10946705221130467

Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., Lu, V. N., Paluch, S., & Martins, A. (2018). Brave new world: Service robots in the frontline. *Journal of Service Management*, *29*(5), 907–931. doi: 10.1108/JOSM-04-2018-0119

Wolf, L., & Steul-Fisher, M. (2023). Factors of customers' channel choice in an omnichannel environment: A systematic literature review. *Management Review Quarterly*, *73*, 1579–1630. doi: 10.1007/s11301-022-00281-w

Wong, N. Y. (2004). The role of culture in the perception of service recovery. *Journal of Business Research*, *57*(9), 957–963. doi: 10.1016/S0148-2963(03)00002-X

World Bank. (2024). *Services, value added (% of gdp).* `https://data.worldbank.org/indicator/NV.SRV.TOTL.ZS?end=2021&name_desc=false&start=1960&view=chart`. (Accessed: 16 January 2025)

Wu, L., Fan, A. A., & Mattila, A. S. (2015). Wearable technology in service delivery processes: The gender-moderated technology objectification effect. *International Journal of Hospitality Management*, *51*, 1–7. doi: 10.1016/j.ijhm.2015.08.010

Wu, S., Fei, H., Qu, L., Ji, W., & Chua, T.-S. (2024). NExt-GPT: Any-to-any multimodal LLM. In *Forty-first International Conference on Machine Learning*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation.* Retrieved from `https://arxiv.org/abs/1609.08144`

Xiao, L., & Kumar, V. (2019). Robotics for customer service: A useful complement or an ultimate substitute? *Journal of Service Research*, *24*(1), 9–29. doi: 10.1177/1094670519878881

Xie, Y., Zhu, F., Wang, J., Liang, R., Zhao, L., & Tang, G. (2018). Long-short term memory for emotional recognition with variable length speech. In *First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)* (pp. 1–4). doi: 10.1109/ACIIAsia.2018.8470341

Xu, B., Li, L., Luo, W., Naseriparsa, M., Zhao, Z., Lin, H., & Xia, F. (2024). Beyond linguistic cues: Fine-grained conversational emotion recognition via belief-desire modelling. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2318–2328). Torino, Italia: ELRA and ICCL.

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., … Chen, E. (2024). Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, *18*(186357). doi: 10.1007/s11704-024-40555-y

Xu, X., & Li, Y. (2016). The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach. *International Journal of Hospitality Management*, *55*, 57–69. doi: 10.1016/j.ijhm.2016.03.003

Xun, J., & Guo, B. (2017). Twitter as customer's eWOM: An empirical study on their impact on firm financial performance. *Internet Research*, *27*(5), 1014–1038. doi: 10.1108/IntR-07-2016-0223

Yang, J., So, J., Zhang, H., Jones, S., Connolly, D. M., Golding, C., … Major, V. J. (2024). Development and evaluation of an artificial intelligence-based workflow for the prioritization of patient portal messages. *JAMIA Open*, *7*(3). doi: 10.1093/jamiaopen/ooae078

Yang, S., & Kruschke, J. K. (2024). An intervention for increasing intention to post online customer reviews. *Journal of Interactive Marketing*, *59*(4), 400–414. doi: 10.1177/10949968241228198

Yang, Y., Chi, M., Bi, X., & Xu, Y. (2024). How does the anthropomorphism of service robots impact employees' role service behavior in the workplace? *International Journal of Hospitality Management*, *122*, 103857. doi: 10.1016/j.ijhm.2024.103857

Yayla-Küllü, H. M., Tansitpong, P., Gnanlet, A., McDermott, C. M., & Durgee, J. F. (2015). Employees' national culture and service quality: An integrative review. *Service Science*, *7*(1), 11–28. doi: 10.1287/serv.2015.0092

Yildirim, H. E., & Iren, D. (2023). Informative speech features based on emotion classes and gender in explainable speech emotion recognition. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (pp. 1–8). doi: 10.1109/ACIIW59127.2023.10388158

Yoon, S., Byun, S., & Jung, K. (2018). Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)* (pp. 112–118). doi: 10.1109/SLT.2018.8639583

Yu, M., Xu, J., & Tang, J. (2024). Managing customer contact centers with delay announcements and automated service. *IISE Transactions*, *56*(2), 115–127. doi: 10.1080/24725854.2023.2183532

Yurtay, Y., Demirci, H., Tiryaki, H., & Altun, T. (2024). Emotion recognition on call center voice data. *Applied Sciences*, *14*(20), 9458. doi: 10.3390/app14209458

Zaki, J., & Williams, W. C. (2013). Interpersonal emotion regulation. *Emotion*, *13*(5), 803–810. doi: 10.1037/a0033839

Zanzotto, F. M. (2019). Viewpoint: Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, *64*, 243–252. doi: 10.1613/jair.1.11345

Zapf, D., Isic, A., Bechtoldt, M., & Blau, P. (2003). What is typical for call centre jobs? Job characteristics, and service interactions in different call centres. *European Journal of Work and Organizational Psychology*, *12*(4), 311–340. doi: 10.1080/13594320344000183

Zhang, C., & Laroche, M. (2020). Brand hate: A multidimensional construct. *Journal of Product & Brand Management*, *30*(3), 392–414. doi: 10.1108/JPBM-11-2018-2103

Zhang, K., Li, Y., Wang, J., Cambria, E., & Li, X. (2022). Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(3), 1034–1047. doi: 10.1109/TCSVT.2021.3072412

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323. doi: 10.1016/j.bspc.2018.08.035

Zhong, J., & Li, W. (2019). *Predicting customer call intent by analyzing phone call transcripts based on CNN for multi-class classification.* Retrieved from `https://arxiv.org/abs/1907.03715`

Zhou, R., Wang, X., Shi, Y., Zhang, R., Zhang, L., & Guo, H. (2019). Measuring e-service quality and its importance to customer satisfaction and loyalty: An empirical study in a telecom setting. *Electronic Commerce Research*, *19*, 477–499. doi: 10.1007/s10660-018-9301-3

Zhou, S., Yi, N., Rashiah, R., Zhao, H., & Mo, Z. (2024). An empirical study on the dark side of service employees' AI awareness: Behavioral responses, emotional mechanisms, and mitigating factors. *Journal of Retailing and Consumer Services*, *79*, 103869. doi: 10.1016/j.jretconser.2024.103869

Zhou, Y., Fei, Z., Yang, J., & Kong, D. (2025). Service with voice: The role of agents' vocal cues in the call center service. *Journal of Business Research*, *192*, 115282. doi: 10.1016/j.jbusres.2025.115282

Zierau, N., Hildebrand, C., Bergner, A., Busquet, F., Schmitt, A., & Leimeister, J. M. (2023). Voice bots on the frontline: Voice-based interfaces enhance flow-like consumer experiences & boost service outcomes. *Journal of the Academy of Marketing Science*, *51*, 823–842. doi: 10.1007/s11747-022-00868-5

Zito, M., Emanuel, F., Molino, M., Cortese, C. G., Ghislieri, C., & Colombo, L. (2018). Turnover intentions in a call center: The role of emotional dissonance, job resources, and job satisfaction. *PLoS One*, *13*(2), 1–16. doi: 10.1371/journal.pone.0192126

Zolfagharian, M., Hasan, F., & Iyer, P. (2018). Customer response to service encounter linguistics. *Journal of Services Marketing*, *32*(5), 530–546. doi: 10.1108/JSM-06-2017-0209

Zorn, S., Jarvis, W., & Bellman, S. (2010). Attitudinal perspectives for predicting churn. *Journal of Research in Interactive Marketing*, *4*(2), 157–169. doi: 10.1108/17505931011051687

Zweig, G., Siohan, O., Saon, G., Ramabhadran, B., Povey, D., Mangu, L., & Kingsbury, B. (2006). Automated quality monitoring in the call center with ASR and maximum entropy. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* (Vol. 1, p. I-I). doi: 10.1109/ICASSP.2006.1660089

Appendices

Appendix A. Statistics of samples in train, validation, and test set

Each set consists of satisfied (SAT) and dissatisfied (DSAT) conversations. Statistics about the audio are in “seconds”.

Statistic	**Train (50%)**			**Validation (25%)**			**Test (25%)**			**Total**
	SAT	DSAT	Total	SAT	DSAT	Total	SAT	DSAT	Total	
Number of conversations	472	100	572	236	50	286	235	51	286	1144
Mean audio length (s)	221.2	262.2	228.4	229.7	249.5	233.1	227.9	268.8	253.2	231.3
SD audio length (s)	150.8	134.6	148.8	125.5	152.3	130.5	124.7	236.5	151.1	144.9
Minimum audio length (s)	63.0	94.9	63.0	62.5	69.1	62.5	23.8	55.1	23.8	23.8
Maximum audio length (s)	1892.1	902.9	1892.1	873.8	658.7	873.8	836.1	1511.5	1511.5	1892.1
Mean number of sentences	91.3	95.2	92.0	97.0	83.4	94.6	95.3	90.0	94.3	93.2
SD number of sentences	42.4	36.8	41.5	47.3	34.0	45.5	42.0	40.3	41.7	42.5
Minimum number of sentences	23	27	23	23	27	23	17	29	17	17
Maximum number of sentences	265	221	265	247	195	247	226	204	226	265

Appendix B. Number of sentences by agent and caller over all sets

Statistic	Agent			Caller			Total		
	SAT	DSAT	Total	SAT	DSAT	Total	SAT	DSAT	Total
Mean number of sentences	51.2	49.1	50.9	42.5	41.9	42.4	93.7	90.9	93.2
SD number of sentences	23.1	21.6	22.8	25.2	21.2	24.5	43.5	37.1	42.5
Minimum number of sentences	10	12	10	8	9	8	17	27	17
Maximum number of sentences	131	118	131	155	105	155	265	221	265

Appendix C. Model structure

All models had a similar architecture, containing an LSTM layer, a linear layer (32 nodes), a linear output layer, and a SoftMax activation function. The model was trained with a batch size of 16 and for a maximum of 150 epochs, with an early stopping patience of 3 epochs. Adam optimization was used, the learning rate was 0.00001, and a cross-entropy loss was implemented. After each training cycle, the model with the best validation performance is saved.

Appendix D. LSTM

An LSTM cell is a gradient-based recurrent neural network that processes data sequentially. The LSTM call was proposed to overcome the vanishing gradient problem. It consists of an input gate 9.1, forget gate 9.2, and output gate 9.3, and the cell state update (Hochreiter & Schmidhuber, 1997)[1].

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{9.1}$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{9.2}$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{9.3}$$

Where:

- σ: sigmoid function
- w_x: weight vector for gate x
- h_{t-1}: output of previous LSTM block at time $t-1$
- x_t: input at time t
- b_x: bias of gate x

[1] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Cell state update:

$$g_t = \tanh(w_c[h_{t-1}, x_t] + b_c) \tag{9.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{9.5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{9.6}$$

Where:

- c_t: cell state at time t
- g_t: candidate cell state at time t
- $\odot$: element-wise multiplication

A **bidirectional LSTM** adds a reverse LSTM layer to process the sequence backwards (Schuster & Paliwal, 1997)[2]. The backward LSTM equations are:

$$i_t^{back} = \sigma(w_i^{back}[h_{t+1}^{back}, x_t] + b_i^{back}) \tag{9.7}$$

$$f_t^{back} = \sigma(w_f^{back}[h_{t+1}^{back}, x_t] + b_f^{back}) \tag{9.8}$$

$$o_t^{back} = \sigma(w_o^{back}[h_{t+1}^{back}, x_t] + b_o^{back}) \tag{9.9}$$

$$g_t^{back} = \tanh(w_c^{back}[h_{t+1}^{back}, x_t] + b_c^{back}) \tag{9.10}$$

$$c_t^{back} = f_t^{back} \odot c_{t+1}^{back} + i_t^{back} \odot g_t^{back} \tag{9.11}$$

$$h_t^{back} = o_t^{back} \odot \tanh(c_t^{back}) \tag{9.12}$$

[2]Schuster, M. & Paliwal, K.K. (1997). Bidirectional recurrent neural networks. In *IEEE Transactions on Signal Processing, 45*(11), 2673–2681. doi: 10.1109/78.650093

Appendix E. Cross-attention

Bi-directional cross-attention combines text and audio features as follows (Gheini, Ren, & May, 2021)[3]

Since the text and audio signals are not the same size, they are first mapped onto the same hidden space:

$$\text{audio} = \text{ReLU}(\text{Linear}_{audio}(\text{audio\_sequences}))$$
$$\text{text} = \text{ReLU}(\text{Linear}_{text}(\text{text\_sequences}))$$

For each element in one signal, the model calculates its interactions with all elements in the other signal. This is done by performing element-wise multiplication and accumulation to calculate the values:

$$\text{weighted}_t[i,j,k] = \sum_{t=1}^{T_{audio}} \sum_{s=1}^{T_{text}} \text{audio}[i,t,s] \cdot \text{text}[j,s,k]$$

A SoftMax activation is applied to the calculated interactions to obtain a set of weights. These weights represent the importance or attention given to each element in the second signal when considering the audio signal:

$$\text{output}_w = \text{softmax}(\text{text})$$

Then, the text signal is re-weighted by combining its original elements based on the obtained weights. This results in a weighted combination that highlights the most relevant parts of the text signal for each element in the audio:

$$\text{weighted}_s[i,j,k] = \text{output}_w[i,j,k] \cdot \text{weighted}_t[i,j,k]$$

Finally, the weighted combination is processed through a bidirectional LSTM layer. This layer helps the model capture temporal dependencies and patterns in the combined information. It is bidirectional, as both previous and future audio features can

[3]Gheini, M., Ren, Xiang & May, J. (2021). Cross-attention is all you need: Adapting pretrained transformers for machine translation. Retrieved from https://arxiv.org/abs/2104.08771 doi: 10.48550/arXiv.2104.08771

be of interest for each text feature:

$$\text{output} = \text{bidirectionalLSTM}(\text{weighted}_s)$$

Appendix F. Graphs: Airport

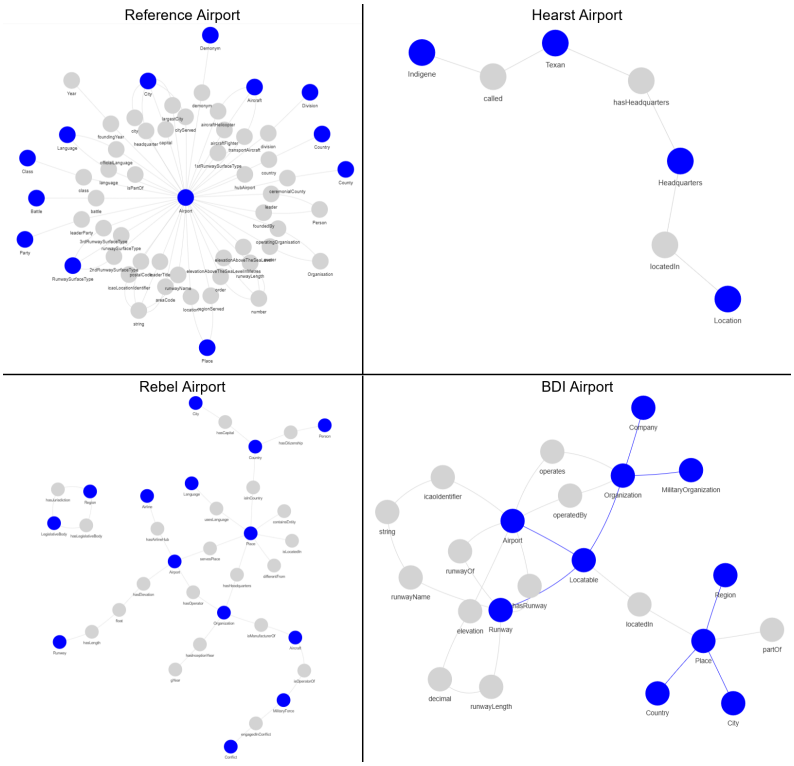

Figure 9.1: Graphs of the airport category.

Note. The blue dots indicate the ontology's concepts (classes), while the grey nodes are supporting elements, such as properties, instances, or domains.

Appendix G. Graphs: Written work

Figure 9.2: Graphs of the written work category.

Note. The green dots indicate the ontology's concepts (classes), while the grey nodes are supporting elements, such as properties, instances, or domains.

Appendix H. Questions: Airport

- What does the dataset reveal about the airports?
- What regions or countries are represented in the dataset?
- How many unique runway lengths are present in the dataset?
- Are there any patterns or standardizations in the naming of runways across the airports in the dataset?
- What types of organizations operate or have a presence at the airports in the dataset?
- Are there any specific runway length patterns or standardizations across the airports in the dataset?
- Who are the primary operators or owners of the airports in the dataset?
- What countries or regions are associated with the airport locations mentioned in the dataset?
- Are there any civilian airports among the airports mentioned in the dataset?
- Are there any military airbases among the airports mentioned in the dataset?
- What is the average runway length for the airports mentioned in the dataset?
- How many distinct runway lengths are there across all the airports in the dataset?
- What naming conventions, if any, are used for runways at different airports in the dataset?
- What is the most common continent or region where the airports are located according to the dataset?
- What countries or regions are represented by the airport locations in the dataset?
- What countries or regions are represented by the airport locations in the dataset?

Appendix I. Questions: Company

- What does the dataset reveal about companies?
- What is the distribution of net income among the companies in the dataset?
- Which industries are represented in the dataset based on the products they produce?
- What is the range and distribution of employee count among the companies in the dataset?
- Are there any patterns or trends in the location distribution of the companies in the dataset?
- What is the extent and nature of subsidiary relationships among the companies in the dataset?
- When were the companies in the dataset established?
- Are there any other parent-child company relationships besides AmeriGas and PHILIPPINE_ENTERTAINMENT_PORTAL_INC.?
- In which countries are most of the companies located in the dataset?
- Can we identify any common product categories among the companies in the dataset?
- Is there a geographical concentration of companies within specific regions or countries?
- How does the employee count vary across different company sizes in the dataset?
- What is the minimum, maximum, and median employee count for all companies in the dataset?
- Who are the parent companies and their subsidiaries for all companies in the dataset?
- What is the standard deviation of net income for all companies in the dataset?
- Are there any other industries present in the dataset besides Energy and Pharmaceutical?
- Are there any correlations between company size and net income in the dataset?
- What is the timeline of establishment for all companies in the dataset?

- What is the range and distribution of employee size for all companies in the dataset?
- What are the primary industries associated with the companies in the dataset?
- Where are the majority of the companies in the dataset located?
- What is the range of revenue and net income among the companies in the dataset?
- What is the average net income and operating income for all companies in the dataset?
- What is the average revenue and net income for all companies in the dataset?
- What is the average net income and revenue range for the companies in the dataset?
- Are there any other industries present in the dataset besides Energy and Drug-Maker?
- What is the distribution of net income among all companies in the dataset?
- What is the distribution of net income among the companies in the dataset?
- Are there any other industries present in the dataset besides Energy and Pharmaceutical?

Appendix J. Questions: Written work

- What does the dataset reveal about the written works?
- Who are the authors associated with the dataset, and what other written works have they produced?
- What other written works have been produced by Diane Duane and Garth Nix?
- What genres do the written works in the dataset belong to, and is there a predominant genre or a mix of genres?
- In which countries are the written works in the dataset set, and is there a pattern in the geographical representation?
- What recurring themes or motifs can be found across the written works in the dataset?
- What is the average and median number of pages for the written works in the dataset, and does this vary significantly between genres or authors?
- What publishers are involved in the publication of the written works in the dataset, and how has the distribution of these publishers changed over time?
- Are there any sequels, prequels, or series connections between the written works in the dataset?
- List all unique authors in the dataset along with their other published works.
- Are there any series connections between the written works in the dataset?
- In which countries were the written works in the dataset originally published?
- Which publishers are associated with the texts in the dataset?
- In which countries are the written works in the dataset set?
- What other written works have been produced by Garth Nix?
- Who are the authors represented in the dataset?
- What other written works have been produced by Diane Duane?
- What is the average number of pages for the written works in the dataset?
- Who are the authors of the written works in the dataset?
- What genres do the works in the dataset belong to?
- What are the genres of the written works in the dataset, and is there a predominant genre?

- What other written works have Diane Duane and Garth Nix produced?
- In which genres do the written works in the dataset belong?
- What other written works have been produced by Garth Nix and Diane Duane?
- Who are the authors associated with the dataset, and what other written works have they produced?
- What genres do the written works in the dataset belong to, and is there a predominant genre or a mix of genres?
- In which countries are the written works in the dataset set, and is there a pattern in the geographical representation?
- In which countries are the written works in the dataset set?
- What recurring themes or motifs can be found across the written works in the dataset?

Abstract

Services shape our daily lives, from ordering coffee and streaming our favorite TV show to navigating technical issues and accessing healthcare. These customer service interactions generate vast amounts of information every day. However, organizations struggle to systematically analyze and learn from these interactions due to their volume and complexity. This dissertation examines how artificial intelligence (AI) can automatically extract meaningful patterns and insights from customer service interactions, transforming them into actionable information for organizations. Central to this approach is the recognition that service interactions contain multiple layers of information: explicit content (what is actually said), implicit signals (emotions, dissatisfaction, and call quality), and broader interaction-level patterns. Each signal requires specialized analytical techniques to unlock its value.

The dissertation contains six interconnected studies, starting with a systematic literature review that establishes how technology can augment human service agents. From a customer-centric perspective, a comparative analysis of neural network approaches for speech emotion recognition shows that simpler models perform as well as more complex ones, while being more efficient to deploy. Building on this audio-based foundation, subsequent research investigates customer dissatisfaction detection from both text and audio, demonstrating that multimodal approaches significantly outperform single-channel methods. Shifting to a service agent-focused perspective, AI can automatically identify different response strategies that companies use in social media service interactions, with custom-trained models outperforming general-purpose pretrained models. The research then explores how computational approaches can assist human evaluators in call quality monitoring while highlighting the continued importance of human judgment. Finally, a curiosity-driven approach for aggregated knowledge extraction demonstrates how AI can dynamically structure

collections of service-related documents into ontologies.

Based on these chapters, this dissertation offers several key insights. First, service interactions contain various layers of information that can be extracted automatically. When aggregated across interactions, this information enables organizations to move from reactive problem-solving for individual interactions to developing strategic insights over multiple interactions. Second, the optimal AI approach varies depending on the specific task and context. The results show that simpler, domain-specific models often outperform complex, general, or pretrained models. Third, the findings suggest that the most effective applications emerge when AI technologies are combined with human expertise, rather than replacing human judgment.

This dissertation establishes customer service interactions as underutilized sources of competitive advantage. It demonstrates that AI can systematically extract value across multiple information layers of service interactions. Rather than automating service delivery, these findings suggest that AI systems can provide analytical support that enhances human understanding of service interactions while preserving the essential human elements of customer service. The studies offer both a theoretical understanding of how AI can process interaction data and practical guidance for organizations that are seeking to implement these technologies effectively, contributing to knowledge at the multidisciplinary intersection of service management, information systems, and artificial intelligence research.

Samenvatting

Dienstverlening vormt ons dagelijks leven, van het bestellen van koffie en het streamen van onze favoriete tv-show tot het navigeren door technische problemen en het verkrijgen van toegang tot gezondheidszorg. Deze klantenservice-interacties leveren dagelijks enorme hoeveelheden waardevolle informatie op. Organisaties hebben echter moeite om deze interacties systematisch te analyseren en ervan te leren vanwege hun volume en complexiteit. Dit proefschrift onderzoekt hoe kunstmatige intelligentie (AI) automatisch betekenisvolle patronen en inzichten kan extraheren uit klantenservice-interacties, en deze kan transformeren in bruikbare informatie voor organisaties. Centraal in deze benadering staat de erkenning dat service-interacties meerdere lagen informatie bevatten: expliciete inhoud (wat daadwerkelijk wordt gezegd), impliciete signalen (emoties, ontevredenheid en gesprekskwaliteit), en bredere patronen op interactieniveau. Elk signaal vereist gespecialiseerde analytische technieken om de waarde ervan te verkrijgen.

De dissertatie bevat zes onderling verbonden studies, beginnend met een systematisch literatuuronderzoek dat vaststelt hoe technologie menselijke service-medewerkers kan ondersteunen. Vanuit een klantgericht perspectief toont een vergelijkende analyse van neurale netwerkaanpakken voor spraakemotieherkenning aan dat eenvoudigere modellen net zo goed presteren als complexere modellen, terwijl ze efficiënter zijn om in te zetten. Voortbouwend op deze op audio gebaseerde basis onderzoekt daaropvolgend onderzoek de detectie van klantontevredenheid vanuit zowel tekst als audio, waarbij wordt aangetoond dat multimodale benaderingen aanzienlijk beter presteren dan single-channel methoden. Met een verschuiving naar een op de service-medewerker gericht perspectief kan AI automatisch verschillende responsstrategieën identificeren die bedrijven gebruiken in social media service-interacties, waarbij op maat getrainde modellen beter presteren dan algemene voor-

getrainde modellen. Het onderzoek verkent vervolgens hoe computationele benaderingen menselijke beoordelaars kunnen ondersteunen bij het monitoren van gesprekskwaliteit, terwijl het blijvende belang van menselijk oordeelsvermogen wordt benadrukt. Tot slot demonstreert een nieuwsgierigheidsgedreven benadering voor geaggregeerde kennisextractie hoe AI dynamisch verzamelingen van servicegerelateerde documenten kan structureren in ontologieën.

Op basis van deze hoofdstukken biedt deze dissertatie verschillende belangrijke inzichten. Ten eerste bevatten service-interacties verschillende lagen informatie die automatisch kunnen worden geëxtraheerd. Wanneer deze informatie wordt geaggregeerd over interacties heen, stelt dit organisaties in staat om van reactief probleemoplossen voor individuele interacties te bewegen naar het ontwikkelen van strategische inzichten over meerdere interacties. Ten tweede varieert de optimale AI-benadering afhankelijk van de specifieke taak en context. De resultaten tonen aan dat eenvoudigere, domeinspecifieke modellen vaak beter presteren dan complexe, algemene of voorgetrainde modellen. Ten derde suggereren de bevindingen dat de meest effectieve toepassingen ontstaan wanneer AI-technologieën worden gecombineerd met menselijke expertise, in plaats van menselijk oordeelsvermogen te vervangen.

Deze dissertatie positioneert klantenservice-interacties als onderbenutte bronnen van concurrentievoordeel. Het toont aan dat AI systematisch waarde kan extraheren over meerdere informatielagen van service-interacties. In plaats van dienstverlening te automatiseren, suggereren deze bevindingen dat AI-systemen analytische ondersteuning kunnen bieden die het menselijk begrip van service-interacties verbetert, terwijl de essentiële menselijke elementen van klantenservice behouden blijven. De studies bieden zowel een theoretisch begrip van hoe AI interactiegegevens kan verwerken als praktische begeleiding voor organisaties die deze technologieën effectief willen implementeren, en dragen bij aan kennis op het multidisciplinaire snijvlak van servicemanagement, informatiesystemen en kunstmatige intelligentie.

Acknowledgments

A PhD is often described as a journey, and now, nearing the end of mine, I understand why. When I began over four years ago, in the middle of the COVID-19 pandemic, I could only imagine the path in front of me. And what a journey it has been. Four years filled with moments of excitement, unexpected challenges, and endless hard work. Along the way, countless ideas and projects developed, some of which found their way into this dissertation, while others remained as valuable learning experiences. The run-up to this book involved hours of reading and writing, as well as numerous meetings where ideas were debated and challenges were tackled. Each of these steps contributed to the shape of this experience, which, like any PhD trajectory, was definitely not a straight line, but rather a messy tangle of different routes. Some routes led to results, while others encountered roadblocks, and many circled back before heading in new directions. Various aspects of the work intersected, sometimes helping, sometimes hindering, but all somehow moving forward in ways I never expected.

But no trajectory exists in isolation. It is shaped by the people who supported me, challenged me, encouraged me, and reminded me that I was not doing this alone.

First, I would like to express my deepest gratitude to my supervisors, Alex and Stefano. From the beginning, you helped me navigate research complexities with patience and wisdom, shaping both my work and my growth as a researcher. When experiments failed or ideas led nowhere, you reminded me that these struggles are part of the process and helped me find my way forward. You encouraged me to think more critically and explore ideas more thoroughly than I would have on my own. The questions you posed challenged my assumptions, opened new perspectives, and made this work stronger at every step. I am grateful for your accessibility, for the countless meetings and discussions, and for engaging with half-formed ideas that

eventually became something meaningful. Beyond the research itself, you prepared me not just to complete this PhD, but to continue growing as a researcher. Thank you for believing in this project and in me, and for helping shape the trajectory that brought me here.

Second, I would also like to thank Hans for his contributions as my copromotor and for his role in guiding this research. His approach of providing focused input at key moments proved invaluable; our periodic meetings offered valuable opportunities for reflection and recalibration. The feedback he provided on the written work helped ensure the research maintained its direction and quality.

I would also like to express my sincere gratitude to the Faculty of Management and the Department of Organization for the support they provided throughout the project. I thereby thank all my colleagues in the department for the shared experiences and academic exchanges we have had. These interactions have greatly enriched my journey and contributed to my growth as a researcher. Thank you for being a part of my academic development.

Further, I would like to thank the Department of Computer Science and the Beta Faculty. Although it was not my primary faculty, the department provided a welcoming and supportive environment throughout my doctoral studies. The opportunities to present my work have contributed greatly to my development as a researcher. I would like to specifically thank Jesse, who is a co-author on the paper in Chapter 7. Additionally, I am grateful for the opportunity to continue my academic career within the department as a postdoctoral researcher.

I want to extend a special shout-out to my fellow PhD students who have been by my side throughout this journey. A big thanks to Leonie, Fatima, Sarah, Hilal, Saeed, Kishan, Karlygash, and Krist for your unwavering support. You have been there through every high and low, celebrating the successes and helping me navigate the challenges. Your academic insights and encouragement have been heplful to my progress. Beyond the academic support, the mental and emotional support you provided was equally important. I am also grateful for all the fun times, especially our PhD outings, which added joy and balance to this demanding journey. Thank you for understanding and sharing in this experience.

I would also like to thank my colleagues in CAROU, Innovating for Resilience, and the graduate school for creating a welcoming research community. The research meetings, informal discussions, and guidance on broader academic topics like publishing and career development provided a valuable perspective and a sense

of community throughout this journey.

Furthermore, I would like to thank DHL, particularly the team in Maastricht. Their help with the data used in multiple studies of my PhD was central for the related papers. Their willingness to engage with the research process and share industry perspectives made this work more practically relevant and impactful.

Then, I would like to thank a few people who offered a personal connection and understanding beyond academic support. Special thanks to Rüveyda, Leonie, Nadja, and Fatima - colleagues, yes, but also so much more. Your special way of connecting and the personal encouragement you have given me have been invaluable in supporting me through this PhD.

I want to express my deepest thanks to my family, who have always been there for me. Allereerst, lieve mama en papa, wil ik jullie van harte bedanken. Jullie hebben altijd in mij geloofd en hebben mij gesteund gedurende mijn eerdere studies en dit PhD-traject. Toen ik in 'het buitenland' wilde studeren, hielpen jullie met alle papierwerk en administratie. Toen ik vroeg of twee masters en een jaartje extra ook oké waren, zeiden jullie ja zonder aarzelen. En toen ik terug thuis kwam wonen om deze PhD te beginnen, verwelkomden jullie mij met open armen. Jullie hebben mij geholpen om dit punt te bereiken, en jullie ongelooflijke steun door alle goede en slechte tijden betekent alles voor mij. Jullie zijn altijd mijn grootste supporters geweest, aan mijn zijde bij elke stap van de weg. Nogmaals, heel erg bedankt. Ik wil ook mijn broers bedanken omdat ze mij altijd hebben aangemoedigd en gesteund.

To my dear partner, Tim, who patiently endured all the long hours, busy schedules, and PhD chaos. Thank you for your unwavering support and for always being convinced that everything would work out, even when I was not so sure about it myself.

Finally, I would like to thank all my friends. Thank you for being there through all the struggles and for listening to all the complaints. But also, thank you for celebrating all the joyous moments with me. Your friendship has been a source of strength and happiness, and I am so grateful to have you in my life. You have made this journey so much brighter.