Supporting Online Customer Feedback Management with Automatic Review Response Generation

Dzmitry Katsiuba
University of Zurich
katsiuba@ifi.uzh.ch

Tannon Kew
University of Zurich
kew@cl.uzh.ch

Mateusz Dolata
University of Zurich
dolata@ifi.uzh.ch

Gerhard Schwabe
University of Zurich
schwabe@ifi.uzh.ch

Abstract

The growing amount of online reviews plays a significant role in a business' image and performance. Businesses in the hospitality industry often lack necessary resources to organize and manage online customer feedback and are therefore likely to search for alternative ways to handle this. AI-based technologies may offer valuable solutions. However, there is currently little research on if and how AI solutions may support the process of responding to online customer feedback in the hospitality industry. This paper presents and evaluates a concept for assisting customer feedback management with automatically generated responses to online reviews. Our solution contributes to ongoing investigations into text generation applications for supporting human authors and also proposes new approaches and potential business models for managing online customer feedback.

1. Introduction

Online opinion-sharing platforms and social media have changed the way customers make purchasing decisions. Customers typically browse Internet resources looking for additional information and feedback from other customers, many of whom share their own experience by leaving reviews on various platforms. As a result, an ever-growing number of opinions that are freely accessible to all internet users has given rise to electronic word-of-mouth (eWOM).

Traditional "offline"-businesses such as hotels and restaurants are among those that have become heavily dependent on eWOM. A large portion of the interaction, before the actual experience (seeking information, planning, booking) and after (sharing, reviewing), happens online. Ratings and reviews about their services and products left on platforms like *Tripadvisor*, *Yelp* or *Booking.com* are used by other customers to form an impression of a business [1]. Therefore, there are few hospitality businesses today that can ignore their online image and still perform competitively.

Consequently, many businesses are aware of the role of online reviews and are interested in online customer feedback management (CFM). By responding to reviews appropriately, businesses can use these online platforms as a direct communication channel to influence customer relationships and public discourse [2]. Good responses can help to improve the online image, acquire new customers, foster customer loyalty, and as a result, increase sales [3, 4].

However, responding to an online review is not a trivial task. A suitable response often requires considerable know-how and has to align with the customer's feedback, addressing its content and sentiment appropriately. Moreover, a review response serves multiple purposes and must consider different target readers: the review author, who is primarily interested in a solution to any issues raised or a reaction from the business; potential customers, who may read the response while making future purchase decisions; and web search engines, which analyze the content to optimize their search results. Additionally, the number of reviews being published is increasing from year to year [5, 6]. Small and medium-sized enterprises (SMEs) often lack the necessary resources and time needed to manage online customer feedback effectively. Many businesses might even lack personnel with an adequate command of language to compose effective responses, especially if they serve multinational guests who write reviews in various languages.

Such challenges drive businesses to consider alternative solutions, including outsourcing parts of the CFM process to third-party companies. For example, collecting, analyzing, summarizing, and even formulating responses to reviews published on various online platforms are all services offered by external feedback management providers. In this context, the quality of outsourced responses depends on the organization of their services and the qualifications of the authors they employ.

The development of natural language processing (NLP) and AI solutions reveals new opportunities in the area of CFM. Text generation can be applied to

URI: https://hdl.handle.net/10125/79357
978-0-9981331-5-7

automatically formulate responses to online feedback, resulting in potential time savings and reduced demand for personnel specializing in CFM (e.g., IT services). However, it remains unclear if the quality of automatically generated responses can reach a suitable level. Thus, potential AI solutions must first be developed and evaluated.

This study aims to examine one possible scenario in which the authoring of review responses is supported through an automatic response generation system. We address the following research questions:

RQ1. How to design a tool that supports the process of responding to online reviews?

RQ2. *What potential new business models for CFM does the integration of AI enable?*

2. Related Work

2.1 Online Customer Feedback Management

Successfully managing customer relationships relies on a decent understanding of the customers and their needs. In the context of increasing digitalization and the importance of the Internet as the dominant communication and sales channel, it is essential to see the Internet as a valuable data source [7]. Modern practices of customer relationship management have changed. The method of transmitting information has transformed and crossed from traditional word of mouth (WOM) [8] bounded by geography [9] to unbounded and permanent eWOM. The effects of this are well documented, and eWOM is known to influence the perceived image of the business or product [10, 11, 12] as well as on the consumer behavior before and after consuming products or services. Therefore, businesses need to adapt their communication strategies.

The connected and informed consumer forces businesses to react quickly [13, 14] and competently. Hospitality businesses now find themselves in a new online reality that places considerable demands on already limited resources. While many businesses still do not engage with online feedback, whether it be due to a lack of skills, time, personnel, or money, some have begun investing in IT support. For these enterprises, who are trying to optimize online CFM processes with eWOM, the question is no longer *whether* their CFM processes need to change but *how* they can change.

To this end, it helps to gather knowledge from the large volumes of feedback data and publish appropriate responses to many online reviews across different platforms. Several providers on the market already recognize this niche of CFM, most of them based in the US. Companies such as *Customer Alliance, TrustYou, MEDALLIA, REVINATE,* and *GuestRevu* offer services that include collecting and analyzing customer feedback, CFM reporting, and conducting customer surveys, among others. The task of actually reading and responding to online reviews usually remains the responsibility of the business, with relatively few third-party CFM providers offering services here. Therefore, there is potential for growth in review response writing services, involving human authors employed to compose high-quality responses on behalf of a business.

2.2 Natural Language Processing in Customer Feedback Management

Given that much of the feedback provided by customers is in the form of unstructured text, NLP techniques, such as named entity recognition (NER) and sentiment analysis, are crucial in order to be able to effectively and efficiently analyze the data. Applications of machine learning in NLP are continuously raising the bar for tasks involving the efficient analysis of unstructured text data, while simultaneously improving the capabilities of text generation systems.

We recognize that such technologies could further facilitate or even replace certain tasks of CFM. For example, text generation techniques similar to those used in translation services, chatbots, and conversational agents could be used to automatically generate responses to online reviews. Chatbots or conversational agents are being increasingly used as direct synchronous customer touchpoints, and they are enjoying a strong interest in technology-based services. NLP technologies combined with human-like design could promise a service that is always available and maintains a high quality that is very close to real customer service [15]. Such novel services are also known as bots-as-a-services [16]. In the context of asynchronous communication involving online reviews and responses, the possibilities of automatic response generation are yet to be fully explored. Both forms of communication (synchronous and asynchronous) deal with written customer requests and run online. The major challenge for automatic response generation is the complexity of the response requirements. Today's conversational agents often lack the necessary social skills and world knowledge required to generate convincing and detailed responses. Thus, a great deal of work has been dedicated to making these systems more empathic and human-like (e.g., [17, 18]) as well as encouraging them to leverage external knowledge (e.g., [19, 20]).

Information Systems (IS) research has recognized the potential of applying NLP-based bots-as-a-service [21, 22, 23]. The research explores the use of such technologies in face-to-face settings [24, 25], synchronous chat [26, 27], as well as asynchronous communication like answering emails [24].

Additionally, there has been significant interest in supporting group processes by means of AI-based agents, e.g., in collaborative writing [28, 29, 30, 31]. All of these scenarios involve interaction between a single user or a very limited, predefined group of users and NLP-based technologies. Research on using NLP for communication with a broader, unspecified population of users is still scarce. It is unclear how automatically generated texts are perceived and valued in public discourse as well as the potential implications: how would businesses, CFM, and communication on online review platforms change if it were possible to automatically generate responses which feel like they were written by the business' owner or employee? To explore those issues, we assume that an optimal automatic response will resemble one generated by a human.

The generated responses have to exhibit empathic and human-like messages indistinguishable from that of a human. Empathic and sentiment-adaptive responses are perceived to be more social and thus elicit higher customer contact satisfaction [15]. To evaluate the human aspects of utterances and the resulting communication, we use the Turing Test, according to which human evaluators try to determine whether the conversation partner is a machine or a human through text-only communication [32].

Besides human-like factors, review response communication should also be effective. As one of the core theories used in the communication studies and linguistics, the Cooperative principle phrased by Paul Grice [33] provides a basis for the design of such communication. It includes four maxims, representing rational principles: maxims of quantity, quality, relation, and manner. These maxims provide a guideline for the design of utterances in terms of length, relevance, clarity, etc., and explain the negative effects that arise from the violation of these maxims [34, 35, 36].

3. Method

This study is a part of a larger design science research (DSR) project "Smart Responses" [37]. The DSR process consists of six iterative steps, starting with problem identification, over design and development, to communication. When conducting our study, we largely follow Peffers' DSR methodology [38]. Thus, in the first step, we describe our motivation (see Section 1), identify the problem, and explain the solution objectives for the field of CFM (see Section 4). Then, based on the objectives and goals, we define the process and artifact designs and subsequently develop the targeted solution (see Section 5). The demonstration was carried out within the project "Smart Responses".

Finally, in order to observe how effective and efficient the implemented tool is, we conduct an evaluation (see Section 6). For this, we survey potential customers and carry interviews with direct application users within the context of the "Smart Responses" project. The collected results help validate the process and component architecture. Furthermore, analysis of the results provides useful insights for potential modifications to the artifact as well as opportunities for new process workflows and business models.

4. Problem identification and solution objectives

The "Smart Responses" project was initiated by a management-owned start-up Feedback Management Provider (FMP), which offers services in feedback management for the hospitality and tourism industry. FMP supports its clients by responding to guest reviews. Currently, they use a software solution developed and maintained by an external IT support that aggregates online reviews and manages responses composed by human authors. This existing solution does not provide any automatization for response composition. The "Smart Responses" team includes two research partners with scientific expertise in computational linguistics (CL) and information management (IM). The project aims to contribute towards a better understanding of the design problem and how responding to reviews can be organized and supported with NLP and AI technologies.

The following scenario describes the position of FMP and the goals of the project. Imagine a medium-sized pizzeria "Roma" that has fewer than 20 employees. They primarily focus on managing the internal processes to provide the best quality pizzas for their customers. Like most businesses in the hospitality sector, they already have a profile on *Tripadvisor* and *Yelp*. They are aware of the importance of guest feedback but do not deal with online reviews because of insufficient resources. This leads to the reviews going unseen and unanswered. Though most reviews are good, a few negative ones that heavily criticize the food quality remain visible for everybody to see.

The pizzeria now has the possibility to outsource CFM to the intermediary FMP, which seeks to ensure that the business' online reputation is not harmed (see Figure 1). FMP employs professional authors, who are familiar with the pizzeria and are tasked with reading the unanswered reviews and quickly formulating a response. Naturally, this response must be personalized and meet the requirements and specifications of the pizzeria "Roma" regarding the type and style of message they would like to publish. This model has proven successful, and, as a result, the number of FMP clients is growing. Since it takes a lot of resources and

administrative effort, the task also poses considerable challenges for FMP. In order to meet all quality requirements and write the best possible and up-to-date responses, the human authors must continuously inform themselves about the businesses they represent, including current special offers, staff, dishes, etc. Furthermore, the responses have to be written within a limited amount of time. This creates a logistical and qualitative challenge for FMP, whose employees struggle to meet the demand in an acceptable time.

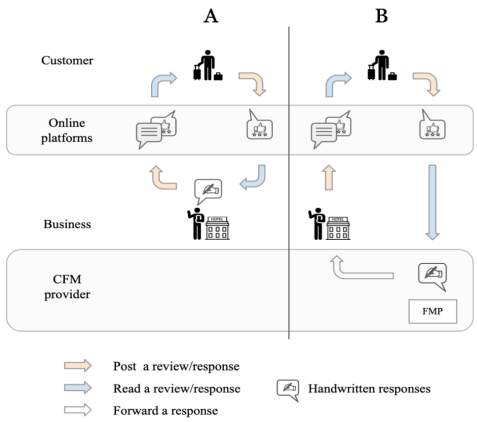

Figure 1. Outsourcing scenario of responding to online reviews

The process of responding to reviews at FMP involves several steps. First, the reviews from different platforms are collected and centralized by FMP. In the next step, review content is analyzed by a human administrator and assigned to one of the authors working for FMP. Subsequently, the authors read the review to be answered and, if necessary, inform themself about the business under consideration or collect further information. Afterwards, the author prepares a response. In the end, the response is validated by the administrator, and if all quality criteria are met, the response is forwarded to the business. In the case of an insufficient response, the response may be edited directly by the administrator or sent back to the author for revision. The final decision about the publication of a response is then made by the business (FMP's client) via FMP's custom software platform. This process involves a socio-technical system in which human actors handle all process steps with no intelligent support systems.

The goal of the "Smart Responses" project is to support the process of responding to reviews by developing a solution objective. We propose an intelligent response generator that is capable of drafting review responses for pizzeria "Roma" and other FMP clients. In this project, our CL partners pursue this solution and develop a tool for this purpose. The IM partner uses the available data and investigates the quality of the generator from the customer's point of view and its impact on the CFM process. Ideally, the pizzeria "Roma" guests do not notice that the answer comes initially from an intelligent tool, which passes the Turing Test [32].

5. Design and Development of Response Generator

In the current study, as part of the "Smart Responses" project (see Figure 3-B), we develop a system that attempts to perform one part of the process: response generation. The aim of such a system is to assist third-party providers that are increasingly engaged by a wide range of businesses due to an inherent lack of resources and know-how. Scenario A in Figure 1 illustrates the typical process when the business manages online customer feedback itself. In the first step of outsourcing (see Figure 3-B), an FMP performs the tasks of responding to reviews. The Pizzeria "Roma" from the "Smart Responses" project is currently in this scenario. By integrating an AI (see Figure 3-C), we step over to the first automation phase, where an intelligent response generator pre-formulates the review responses.

The proposed AI solution is an automatic response generation system, which is entirely data-driven and comprises three main components: (i) a preprocessing module, (ii) a generation model, and (iii) a postprocessing module. Figure 4 illustrates the overall system architecture.

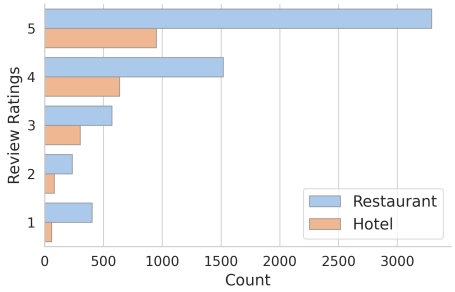

Figure 2. Distribution of review ratings in the training dataset

Data: In order to learn the task of review response generation, we compile a dataset of approximately 8000 hospitality review-response pairs written by a team of FMP authors. Of the total 8000 examples, 75% are in the restaurant domain, and the remaining 25% are for hotels. Review ratings range from 5 (strongly positive) to 1 (strongly negative). Figure 2 shows the distribution of review types in our dataset. The vast majority of reviews are positive (POS; i.e., ratings 4-5), which poses

further challenges for learning automatic response generation for neutral and negative reviews (NEG; i.e., ratings 1-3) as minority classes. For efficiency, we randomly sample roughly 400 pairs for testing and validation splits and use the remainder for training.

Preprocessing: Review responses typically follow a formulaic structure consisting of a greeting, body, and a salutation. Since greetings and salutations are generally standardized phrases and potentially customizable towards specific businesses, we focus only on generating the body of the response and replace known greeting and salutation phrases with placeholders. Additionally, we use the spaCy NLP toolkit to mask out any identifiable information such as personal names, email addresses, and phone numbers, replacing these with special tokens.

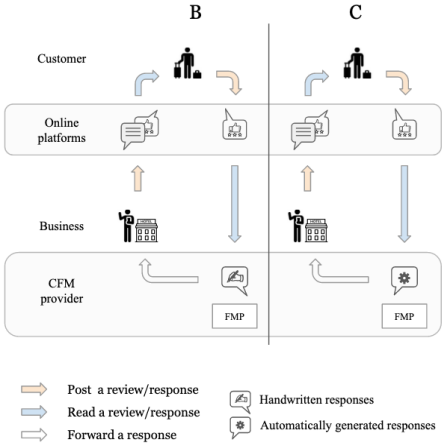

Figure 3. Solution architecture

Generation Model: In order to automatically generate a response text to a customer review, we frame the task as a sequence-to-sequence (seq2seq) modeling task [39, 40]. Given an input review as a sequence of tokens $X = (x_1, \ldots, x_n)$, we aim to derive an informative encoding Z which is used to condition the auto-regressive generation of a target response text $Y = (y_1, \ldots, y_m)$.

Our tool leverages recently proposed techniques in seq2seq pretraining and transfer learning. We use BART [41], a denoising auto-encoding model based on the popular encoder-decoder transformer architecture (see Figure 4). BART is initially pretrained with a general-purpose objective. As input, it takes a sequence of tokens, which is corrupted with a noise transformation (e.g., token masking, sentence permutation, etc.). The model is trained to reconstruct the original input sequence from its corrupted form. Following this pretraining step, the model can be fine-tuned on task-specific source-target pairs using regular cross-entropy loss. BART has been shown to perform well in seq2seq tasks such as abstractive text summarization and question answering. For our experiments, we opt for the smaller BART-base model, which consists of 6 layers in both the encoder and decoder, and fine-tune it on our high-quality review-response pairs until validation loss stops improving.[1]

To help guide the model in generating responses that are appropriate for the given domain and review rating, we prepend discrete token labels to the input review text. This simple approach has been shown to be effective for guiding text production in generation models [43, 44, 45] and ensures that the encoded representation contains at least some relevant conditioning information even in the case of potentially ambiguous or short reviews. Formally, the target sequence Y is modeled under the parameters θ as

$$Y = \sum_{i=1}^{m} P_\theta(y_i \mid d; r; X),$$

where $d \in D$, the closed set of possible domain values, and $r \in R$, the closed set of possible rating values.

Model Inference and Postprocessing: Following Lewis et al. [41], we generate responses using regular beam search decoding [46] with a beam size of 5 to balance between performance and memory consumption during inference. Since beam search is deterministic, it ensures that our results are reproducible and comprise high probability output sequences for a given input. Once a response has been generated, we replace the placeholder tokens with appropriate phrases and named entities for a given review-response pair. These include personalized aspects such as signature greetings and salutations provided by a business and the reviewer's username, the manager's name, and their contact details, among others.

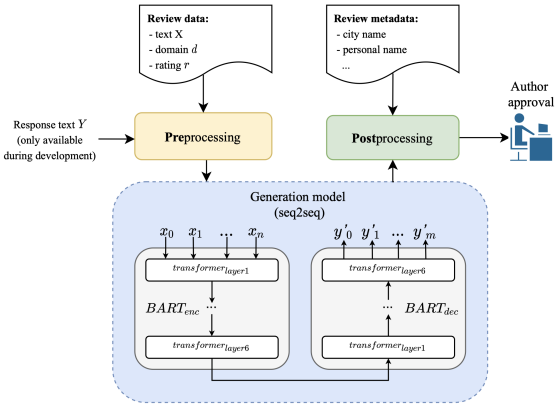

Figure 4. Response generator architecture

[1] We use the Fairseq [42] framework and fine-tune with default settings and an effective batch size of 4096 tokens. Fine-tuning runs for 20 epochs and is performed on a single NVIDIA GTX TITAN X GPU, taking approximately 8 hours.

6. Evaluation

6.1. Evaluation Methods

At the evaluation step, we verify whether our system is able to support the process of answering online reviews and provide acceptable responses to readers. To assess the quality of generated responses from a customer perspective, we conducted a survey with students acting as subjects. For this survey, we used a length criterion to sample 100 review-response pairs from the initially randomly sampled test set. Each selected data point contained a review between 42 and 105 words long (mean=76.7; SD=20.8). This allowed us to examine the contextualization features of the generator while controlling both the reading time and the time required to complete the entire survey. The review includes the text written by the customer as well as metadata such as the name of the author, business, city, and the customer rating. All responses were automatically generated. The output from the response generator underwent post-processing, which included the insertion of a standard greeting with the author´s name and the review's star rating if the generated response refers to it.

Each participant was presented with ten review-response pairs. For each pair, subjects were asked to indicate (1) whether they thought the answer was written by a computer or by the company's manager, and (2) whether the response was appropriate for the review in terms of emotion. In Addition, the subjects were also asked a set of questions considering the quality of reviews *and* the responses, respectively (see Table 2), covering all four of Grice's Maxims [33]. While the first two questions offered only two answer options each (see Table 1), the last set of questions were evaluated on a seven-point Likert scale (from 0 to 6, where 0 indicates a highly negative answer and 6 a highly positive or optimal value of a quality feature).

The questionnaire was conducted in English with master-level students of a computer science course, also taught in English. It was stored on the university server and was freely accessible online during the test session. Our choice of participants ensured that they were comfortable with computers and active internet users. The survey was assigned as part of homework. Participation was not mandatory but was highly recommended as an entry point to a larger homework assignment–an essay about quality of online reviews and responses. The survey was designed to take approximately 20-25 minutes. To reduce the time needed to fill out the survey and keep answers as confidential as possible, we decided not to collect demographic information.

In total, we received responses from 30 participants. The review-response pairs were presented to the participants randomly. Because of the number of participants and the randomization procedure, not all of the 100 prepared review-response pairs received the same number of ratings. We obtained in total 246 answers to 84 review-response pairs. The number of evaluations per pair varied: 34 pairs had answers from only two participants each, 50 pairs received more than two ratings (median=3, mean=3.54). For analysis, we considered all pairs with evaluations from more than two participants. Among the 50 responses taken: 10 responses addressed negative reviews (NEG), and the remaining 40 responses covered positive reviews (POS).

For the qualitative Likert scale questions, we report the mean value of the scores received. For the remaining two questions, we used the opinion of the majority: if two subjects stated that the response was written by a computer and one person said the manager, we report the former, i.e., "by a computer."

6.2. Evaluation results

The first evaluation question about the origin of the response was one of the essential considerations in deciding whether the process of responding to reviews had potential for AI support. In 52% of review-response pairs, subjects stated a belief that the response was written by the company's manager (see Table 1). This is interesting given that all responses were automatically generated, and we did not preselect the responses to be of a certain quality.

Table 1. Evaluation results: Response origin and Context match

Question	NEG	POS	Total (%)
The response was written by ...			
- the computer	5	19	24 (48%)
- the manager of the company	5	21	26 (52%)
Does the response fit the context in terms of emotion?			
- It does not fit the context	1	5	6 (12%)
- It fits the context	9	35	44 (88%)

The vast majority of responses generated are reasonable in terms of domain and emotional appropriateness to the input review. For 44 pairs, respondents indicated a good match to review context, while only 12% of responses failed to correspond with the sentiment expressed in the review.

A detailed analysis of the quality of the reviews and responses (see Table 2) shows that most of the reviews were rated as informative and helpful by the subjects. This confirms our intention to select useful reviews for the evaluation.

Evaluation of responses includes features related to language fluency, style, response length, and content. The best scores were achieved by the responses in the categories of fluency and length. The subjects found the language understandable (mean=4.9; SD=0.52) and fluent (mean=4.8; SD=0.64). The length of responses tended to be well perceived by most participants (mean=4.4; SD=0.72). This may be due to the fact that guests do not appreciate answers that are too short or too long, and they expect the length of the answer to be appropriate to the content of the review.

Table 2. Evaluation results: Review and response quality (Mean (SD))

Question	NEG	POS	Total
How informative is the review?	3.9 (0.47)	4.2 (0.88)	4.2 (0.82)
How helpful is the review?	4.1 (0.56)	4.3 (0.90)	4.3 (0.84)
How informative is the response?	3.5 (0.95)	3.4 (1.12)	3.4 (1.08)
How helpful is the response?	2.9 (1.07)	3.0 (1.09)	3.0 (1.08)
How do you find the length of the response?	4.5 (0.48)	4.4 (0.78)	4.4 (0.72)
How understandable is the language of the response?	5.0 (0.42)	4.9 (0.55)	4.9 (0.52)
How fluent is the language of the response?	4.7 (0.53)	4.9 (0.67)	4.8 (0.64)
How individualized is the response?	2.8 (1.35)	2.9 (1.48)	2.9 (1.44)
How specific is the response text to the review test, i.e., does it address points raised in the review?	2.9 (1.55)	3.0 (1.30)	3.0 (1.33)
Is the response text appropriate for the review text, i.e., does it fit the domain (hotel vs. restaurant)?	4.4 (0.65)	4.4 (0.92)	4.4 (0.86)
How do you find the overall quality of the response?	3.2 (1.00)	3.7 (0.90)	3.6 (0.92)

The poorest performance according to our survey subjects was in the content category. Respondents were rather negative about the degree of individualization (mean=2.9; SD=1.44) and the relation between the review content and the response. Response readers were aware of the information mentioned (mean=3.4; SD=1.08) but found it less helpful (mean=3.0; SD=1.08). In general, respondents found the reviews more informative and helpful than the responses. Nevertheless, the general quality of the responses was rated relatively high on average (mean=3.6; SD=0.92).

Besides the overall analysis of the quality characteristics, Table 2 lists the mean values for each of the review rating groups. Comparison of the means of these two groups revealed no significant differences.

7. Discussion

The Matryoshka principle in CFM: A successful review response produced by a third-party CFM provider should–ideally–not reveal its true origin. Instead, businesses that employ these services should benefit by appearing more capable and engaged. Like a Russian Matryoshka doll, the true authorship remains hidden to the public.[2] Our proposed solution introduces a further level of hidden authorship on the side of the third-party CFM provider.

The Matryoshka principle is not entirely unique to CFM. Deployment of bots-as-a-service has gained significant popularity, typically aiming to uphold a level of service expected for human-to-human services without having to employ human resources [21, 22, 23, 27]. Previously, service hotlines or chats were outsourced to dedicated hotline service providers, while now they are increasingly being handled by various kinds of bots. Still, the user is supposed to believe that they are interacting with a genuine business representative. This has led to intensified research on smooth or seamless handovers from human agents to bots [26] and how the perception of interacting with a bot develops [27].

Whether in the case of outsourcing or automating, both scenarios involve a complex concealment process where users are not supposed to know who the true author is. Businesses that engage these services reinforce this concealment and aim to maintain a human-level quality of service facade. This might be motivated by a misconception about users who would expect human contact. With growing demand, however, it is not at all surprising that many businesses consider AI support for their processes. Nevertheless, we claim it is worth questioning the concealment practice. Given that many users can sense that machines are at work in a service context [21, 22, 24], honesty and transparency about who does what might be valued more than tactics

[2] A set of matryoshkas consists of a wooden figure that opens to reveal a smaller figure of the same kind nested inside, which in turn nests another figure, and so on.

following the Matryoshka principle. We call for intensifying research that tests how users react to disclosure. In our context, disclosure could be in the form of a footnote saying, "this response was composed by a computer on behalf of the restaurant owner" or "this response was composed by a computer but approved by the restaurant manager". This could help prevent a user's discontent with a business' responses if automated response generation proliferates.

Quality of generated responses: Looking at the role of AI in CFM and, in particular, the review response process, we can distinguish between two types of potential integration. The first involves the AI performing and completing individual tasks within the process. The other involves a form of facilitative AI used to support or supplement tasks that are still primarily performed by humans.

In order to leverage AI for task completion, the AI must not only be capable of passing the Turing Test [32] but also fulfilling *all* of the requirements placed on human authors. According to the results of our survey, the proposed solution may only serve as a facilitative AI. In terms of process organization, our tool meets the requirements for generating responses quickly [13, 14] for a large volume of reviews [5, 6]. In terms of linguistic fluency, the quality of the generated response is already on a good level. In all cases, the answers were easily understandable and grammatically accurate. Nevertheless, it is clear that the applied NLP technologies are not yet capable of responding to reviews entirely independently since non-human authorship is often detectable. The quality aspect scores from our human evaluation also indicate that there is considerable room for improvement in automatic review response generation in regards to all of Grice's maxims of communication [33]. On the basis of these results, we conclude that our solution may provide valuable support to human authors for the review response CFM task.

Possible improvements and business models: In order to improve the quality of responses and AI integration, we consider three options: (i) technical improvements on the system level, (ii) process improvements involving how the system is exposed to a human author through a well-designed interface, and (iii) a combined solution with the introduction of an additional intelligent solution.

Technical improvements: A major vulnerability revealed by our evaluation is that generated responses are typically not individualized or specific to the input review. We conjecture that this is largely due to a lack of consistent alignments between source and target pairs in the training data, making this a challenging task for seq2seq approaches. The degree to and the manner in which response authors address topics raised in reviews differs considerably between examples. Since the generation model essentially aims to recover the *mode* of the distribution of possible output sequences [47], it effectively learns how to produce the most probable and 'safest' responses, which are typically generic and fail to address specific topics directly. To counter this and to produce more interesting responses, stochastic decoding techniques such as those discussed by Holtzman [48] may be useful.

Furthermore, the model tends to 'hallucinate', generating content that is unsupported by the source text. This is especially observable when the review context is limited. In order to combat this, grounding knowledge could be used as an additional input to the response generation model [19, 49]. This could be drawn from a maintainable and dynamic knowledge source containing up-to-date facts about a specific business entity or customer visit. Lastly, our model is fine-tuned on only a small number of examples. Additional training data would further help the model generalize to unseen contexts and also extend the model to new domains.

Process improvements: Automatically generated responses may be improved through the involvement of a human editor. From this in-process perspective, the solution lies in the smooth integration of the AI tool into the socio-technical system of the FMP. The employees of an FMP complete many tasks before and while they write a response. An intelligent system may also be able to execute some of these (sub)tasks more effectively (e.g., automatic information retrieval based on keywords in a review text), benefiting from some level of human input and resulting in different human-AI interaction scenarios. Developing reliable response generators would naturally broaden the spectrum of possible business models (see Figure 5).

All responses created by the FMP undergo quality assurance checks. In this step, the results of the pre-generated answers can be evaluated and adjusted by the human authors so that all quality characteristics are fulfilled (see Figure 5-D). Through the improved quality of the responses, we see other CFM scenarios in which the steps involving quality assurance become less critical. This would allow responses to be forwarded directly to the businesses, where they are verified by a human and officially published (see Figure 5-C). In these scenarios, the NLP module acts as a co-writer in the process of developing a text [28, 29, 30, 31]. As exemplified here, there are various configurations for how NLP can contribute and how it complements the human author. Ideally, the process reaches the highest level of automation when the responses are published directly to the online platforms (see Figure 5-E). However, the results we obtained indicate that machines still lack abilities to be a stand-alone author at eye level. We are in need of identifying effective and ineffective

configurations of human and NLP-based actors in co-writing settings.

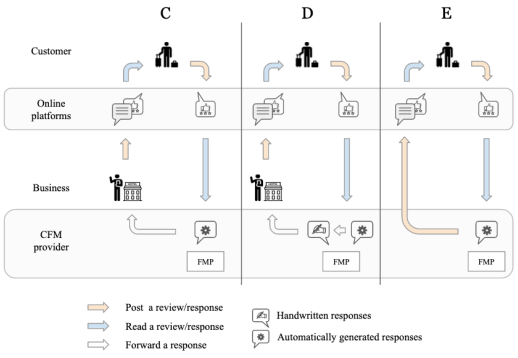

Figure 5. Possible scenarios of responding to online reviews

Combined solution: Following the human-in-the-loop scenario, a mixed solution can also be conceived as another alternative. Here, the FMP's socio-technical system can be further extended by more 'intelligent' software that aims to validate the quality of generated responses in order to support the work of humans better. For instance, identifying, highlighting, and providing suggestions for aspects of the generated response that could be considered vulnerable or erroneous (e.g., hallucinations) would help authors in quickly directing their focus and editing skills towards just these parts of the text. Additionally, combining multiple model outputs with simple drag and drop interfaces may also allow authors to compose high-quality responses based on the model's suggestions quickly. In this way, the aim is not to achieve a perfect automatically generated response through a single model, but to provide effective support to the authors through multiple steps.

8. Conclusion and Future Work

This research explores the use of text generation techniques from NLP for responding to online reviews. The results show that the generated responses are considered grammatically correct and internally coherent but might lack precision and semantic connections to the review. Nevertheless, in more than 50% of cases, the readers attributed authorship to a human. This suggests that NLP can be effectively used for supporting the composition of responses, but they might require improvement by a human author. This has practical implications for CFM providers, who heavily rely on human authors, and information provided by the businesses to provide better answers to guests more quickly. At the same time, the research points to the risk of the Matryoshka principle, which–if it proliferates further–might ultimately reduce the value of a response such that guests or potential guests will no longer pay attention to it.

With our study we contribute transferable design knowledge, which can be used as a basis for the development of new products. We describe a solution architecture for CFM with external support. We have provided a response generator architecture and evaluated the results of an implemented instance.

Future Work: This study does not come without limitations typical of an early design exploration. To enhance the external validity of the study, we intend to increase the number and diversity of participants for the evaluation survey, as well as expanding our automatic response generator to languages beyond English (e.g., German, Italian, etc.).

The composition of the training and test dataset can influence the generated results. Therefore, we additionally plan to focus on the distribution of review sentiments in our future evaluations. This will help us explore the possible applications of our solution and how it depends on the review rating.

Finally, integrating intelligent response generators into the overall socio-technical system in FMP poses challenging questions for the design of human-AI collaboration. To explore this, we plan to conduct an extensive study involving professional response authors as well as novice users. Evaluating responses of the whole system rather than just the NLP module could provide further insights into the perception of responses, their quality, and their value.

Acknowledgment

This study was performed as part of the ReAdvisor project (referred to as "Smart Responses" in the text), an innovation project funded by the Swiss Innovation Agency Innosuisse (project number 38943.1 IP-ICT). ReAdvisor is a collaborative effort of re:spondelligent GmbH (referred to as Feedback Management Provider (FMP) in the text), Welante AG, the Department of Computational Linguistics and the Department of Informatics at the University of Zurich. We thank all project members for their feedback and involvement during the development and the evaluation of the solution. We also thank the anonymous participants of our study, as well as the review team for their valuable advice concerning this manuscript.

References

[1] Sparks, B.A., K.K.F. So, and G.L. Bradley, "Responding to negative online reviews: The effects of hotel responses on customer inferences of trust and concern", Tour. Manag. 53, 2016, pp. 74–85.

[2] Li, C., G. Cui, and L. Peng, "The signaling effect of management response in engaging customers: A study of the hotel industry", Tour. Manag. 62, 2017, pp. 42–53.
[3] Chevalier, J.A., and D. Mayzlin, "The Effect of Word of Mouth on Sales: Online Book Reviews", J. Mark. Res. 43(3), 2006, pp. 345–354.
[4] Ye, Q., R. Law, and B. Gu, "The impact of online user reviews on hotel room sales", Int. J. Hosp. Manag. 28(1), 2009, pp. 180–182.
[5] Statista, "Tripadvisor: number of reviews", Statista, 2020. https://www.statista.com/statistics/684862/tripadvisor-number-of-reviews/
[6] Melián-González, S., J. Bulchand-Gidumal, and B. González López-Valcárcel, "Online Customer Reviews of Hotels: As Participation Increases, Better Evaluation Is Obtained", Cornell Hosp. Q. 54(3), 2013, pp. 274–283.
[7] Rentzmann, R., H. Hippner, F. Hesse, and K.D. Wilde, "IT-Unterstützung durch CRM-Systeme", In H. Hippner, B. Hubrich and K.D. Wilde, eds., Grundlagen des CRM: Strategie, Geschäftsprozesse und IT-Unterstützung. Gabler, Wiesbaden, 2011, 129–155.
[8] Hennig-Thurau, T., K.P. Gwinner, G. Walsh, and D.D. Gremler, "Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?", J. Interact. Mark. 18(1), 2004, pp. 38–52.
[9] Jalilvand, M.R., S.S. Esfahani, and N. Samiei, "Electronic word-of-mouth: Challenges and opportunities", Procedia Comput. Sci. 3, 2011, pp. 42–46.
[10] Hallmann, K., A. Zehrer, and S. Müller, "Perceived Destination Image: An Image Model for a Winter Sports Destination and Its Effect on Intention to Revisit", J. Travel Res. 54(1), 2015, pp. 94–106.
[11] Liu, C.-H.S., and T. Lee, "Service quality and price perception of service: Influence on word-of-mouth and revisit intention", J. Air Transp. Manag. 52, 2016, pp. 42–54.
[12] Pratminingsih, S., "Roles of Motivation and Destination Image in Predicting Tourist Revisit Intention: A Case of Bandung – Indonesia", Int. J. Innov. Manag. Technol. 5(1), 2014.
[13] Xie, K.L., K.K.F. So, and W. Wang, "Joint effects of management responses and online reviews on hotel financial performance: A data-analytics approach", Int. J. Hosp. Manag. 62, 2017, pp. 101–110.
[14] Mattila, A.S., and D.J. Mount, "The impact of selected customer characteristics and response time on E-complaint satisfaction and return intent", Int. J. Hosp. Manag. 22(2), 2003, pp. 135–145.
[15] Diederich, S., M. Janssen-Müller, A.B. Brendel, and S. Morana, "Emulating Empathetic Behavior in Online Service Encounters with Sentiment-Adaptive Responses: Insights from an Experiment with a Conversational Agent", Int. Conf. Inf. Syst., (2019).
[16] Gentsch, P., Künstliche Intelligenz für Sales, Marketing und Service: Mit AI und Bots zu einem Algorithmic Business–Konzepte, Technologien und Best Practices, Springer, 2017.
[17] Harrison, V., L. Reed, S. Oraby, and M. Walker, "Maximizing Stylistic Control and Semantic Accuracy in NLG: Personality Variation and Discourse Contrast", ArXiv190709527 Cs, 2019.
[18] Song, H., Y. Wang, W.-N. Zhang, X. Liu, and T. Liu, "Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation", ArXiv200407672 Cs, 2020.
[19] Gao, C., W. Zhou, X. Xia, D. Lo, Q. Xie, and M.R. Lyu, "Automating App Review Response Generation Based on Contextual Knowledge", ArXiv201006301 Cs, 2020.
[20] Zhao, L., K. Song, C. Sun, Q. Zhang, X. Huang, and X. Liu, "Review Response Generation in E-Commerce Platforms with External Product Information", WWW, ACM (2019), 2425–2435.
[21] Gnewuch, U., S. Morana, M. Adam, and A. Maedche, "Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction", Res. Pap., 2018.
[22] Gnewuch, U., S. Morana, and A. Maedche, "Towards Designing Cooperative and Social Conversational Agents for Customer Service", ICIS, (2017).
[23] Semmann, M., C. Grotherr, P. Vogel, E. Bittner, C. Biemann, and T. Böhmann, "Intelligent Collaboration of Humans and Language-Based Assistants (INSTANT)", ICIS, (2018).
[24] Dolata, M., M. Kilic, and G. Schwabe, "When a computer speaks institutional talk: Exploring challenges and potentials of virtual assistants in face-to-face advisory services", HICSS, (2019).
[25] Färber, A., N. Zigan, M. Dolata, P. Stalder, A.L. Koppitz, and G. Schwabe, "The digital transformation of physician-patient consultations : identifying problems and approaches to improve adherence", HICSS, (2019), 4125–4134.
[26] Poser, M., S. Singh, and E. Bittner, "Hybrid Service Recovery: Design for Seamless Inquiry Handovers between Conversational Agents and Human Service Agents", HICSS, (2021).
[27] Feine, J., S. Morana, and U. Gnewuch, "Measuring Service Encounter Satisfaction with Customer Service Chatbots using Sentiment Analysis", Int. Conf. Wirtsch., (2019).
[28] Wiethof, C., N. Tavanapour, and E. Bittner, "Design and Evaluation of a Collaborative Writing Process with Gamification Elements", ECIS, (2020).
[29] Wiethof, C., N. Tavanapour, and E. Bittner, "Designing and Evaluating a Collaborative Writing Process with Gamification Elements: Toward a Framework for Gamifying Collaboration Processes", ECIS, (2021), 38–61.
[30] Wiethof, C., N. Tavanapour, and E. Bittner, "Implementing an Intelligent Collaborative Agent as Teammate in Collaborative Writing: toward a Synergy of Humans and AI", HICSS, (2021).
[31] Seeber, I., E. Bittner, R.O. Briggs, et al., "Machines as teammates: A collaboration research agenda", Proc. 51st Hawaii Int. Conf. Syst. Sci., (2018).
[32] Turing, A.M., "Computing machinery and intelligence", Mind 54(263), 1950, pp. 433.
[33] Grice, H.P., "Logic and Conversation", In P. Cole and J.L. Morgan, eds., Speech Acts. BRILL, 1975, 41–58.

[34] Jacquet, B., J. Baratgin, and F. Jamet, "The Gricean Maxims of Quantity and of Relation in the Turing Test", 2018 11th Int. Conf. Hum. Syst. Interact. HSI, (2018), 332–338.
[35] Jacquet, B., A. Hullin, J. Baratgin, and F. Jamet, "The Impact of the Gricean Maxims of Quality, Quantity and Manner in Chatbots", 2019 Int. Conf. Inf. Digit. Technol. IDT, (2019), 180–189.
[36] Alba Juez, L., "Verbal irony and the Maxims of Grice's cooperative principle", 1995.
[37] Hevner, A.R., S.T. March, J. Park, and S. Ram, "Design Science in Information Systems Research", MIS Q. 28(1), 2004, pp. 75–105.
[38] Peffers, K., T. Tuunanen, M.A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research", J. Manag. Inf. Syst. 24(3), 2007, pp. 45–77.
[39] Sutskever, I., O. Vinyals, and Q.V. Le, "Sequence to Sequence Learning with Neural Networks", ArXiv14093215 Cs, 2014.
[40] Vaswani, A., N. Shazeer, N. Parmar, et al., "Attention Is All You Need", Conf. Neural Inf. Process. Syst., (2017).
[41] Lewis, M., Y. Liu, N. Goyal, et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension", ArXiv191013461 Cs Stat, 2019.
[42] Ott, M., S. Edunov, A. Baevski, et al., "fairseq: A Fast, Extensible Toolkit for Sequence Modeling", ArXiv190401038 Cs, 2019.
[43] Johnson, M., M. Schuster, Q.V. Le, et al., "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation", Trans. Assoc. Comput. Linguist. 5, 2017, pp. 339–351.
[44] Keskar, N.S., B. McCann, L.R. Varshney, C. Xiong, and R. Socher, "CTRL: A Conditional Transformer Language Model for Controllable Generation", ArXiv190905858 Cs, 2019.
[45] Kumar, V., A. Choudhary, and E. Cho, "Data Augmentation using Pre-trained Transformer Models", ArXiv200302245 Cs, 2021.
[46] Graves, A., "Sequence Transduction with Recurrent Neural Networks", ArXiv12113711 Cs Stat, 2012.
[47] Müller, M., and R. Sennrich, "Understanding the Properties of Minimum Bayes Risk Decoding in Neural Machine Translation", ArXiv 210508504, 2021.
[48] Holtzman, A., J. Buys, L. Du, M. Forbes, and Y. Choi, "The Curious Case of Neural Text Degeneration", Int. Conf. Learn. Represent., (2019).
[49] Prabhumoye, S., K. Hashimoto, Y. Zhou, A.W. Black, and R. Salakhutdinov, "Focused Attention Improves Document-Grounded Generation", ArXiv 210412714, 2021.