She? The Role of Perceived Agent Gender in Social Media Customer Service

Junyuan Ke
University of Rochester
Junyuan.Ke@Simon.Rochester.edu

Yang Gao
University of Illinois Urbana-Champaign
ygao1@Illinois.edu

Huaxia Rui
University of Rochester
Huaxia.Rui@Simon.Rochester.edu

Shujing Sun
The University of Texas at Dallas
Shujing.Sun@UTDallas.edu

Abstract

This work investigated the role of perceived agent gender in customer behavior using a unique dataset from Southwest Airlines' Twitter account. We inferred agent gender based on the first names provided by agents when responding to customers. We measured customer behavior using three outcomes: whether a customer decided to continue the service conversation upon receiving an agent's initial response as well as the valence and arousal levels in their second tweet if the customer chose to continue the interaction. Our identification strategy relied on the Backdoor Criterion and hinged on the assumption that customer service requests are assigned to the next available agent, independent of agent gender. The findings revealed that customers were more likely to continue interactions with female agents than male agents and they were more negative in valence but less intense in arousal with the former group than with the latter.

Keywords: Gender, Stereotype, Customer Service, Social Media

1. Introduction

Imagine that you complained to an airline's Twitter account about a recent flight experience and received a tweet in response from a customer service agent named Andrew. How would you carry on the conversation? Now imagine that, instead of Andrew, the response was signed with the name Alice. Would your reaction be somewhat different simply because of the genders implied by these names? Would you be more or less likely to continue the conversation? If you decided to continue, would your next message be more negative or perhaps less intense?

These are not just intriguing questions for scientifically-minded researchers; they are also in the minds of real-life practitioners. For example, to understand how gender affects support interactions, Olark, a live chat software vendor, conducted a gender experiment[1] in which Sarah Betts, a customer service veteran, changed her name to Samuel and used a male portrait in her profile during one month of live chatting with customers. Even with this rudimentary design, the company was able to identify different behavioral patterns when customers interacted with agents of different perceived genders. In particular, the experiment revealed that customers were more satisfied with yet more abusive towards Sarah than Samuel.

Of course, these analyses and findings need to be rigorously evaluated and carefully interpreted, but why should we expect gender perceptions to influence a customer's interaction with an agent? These influences can be explained by at least two mechanisms rooted in theories and evidence from psychology. The first mechanism involves perceived gender differences in empathy and helping behaviors. Women are generally perceived as more empathetic than men (Hoffman, 1977) and often rated as more helpful, kind, and compassionate than men (Bem, 1974). The second mechanism concerns occupational stereotype, the well-known tendency to view an occupation dominated by one gender as better suited to the characteristics of that gender (Basow, 1992). In the customer service industry, women have historically dominated men in terms of number (Steiger & Wardell, 1995). These two mechanisms may lead to a natural preference for female agents, which could result in a higher probability that customers will continue interacting with such agents.

Moreover, subsequent messages from a customer may be subtly influenced by perceived gender differences. For example, some customers may strategically exaggerate or emphasize their negative experiences in messages to female agents to obtain better redress due to the expectation that women are more compliant and softhearted than men (Eagly, 1983). In addition, customers may be more willing to

[1] See https://blog.olark.com/live-chat-gender-equality-experiment.

URI: https://hdl.handle.net/10125/106604
978-0-9981331-7-1

share their distress with women than men due to the perception that the former is more likely to resonate with and provide emotional support. In either case, female agents may receive messages with more negative valence. Meanwhile, the public nature[2] of social media customer service and cultural expectations regarding the manner toward and treatment of women may regulate the emotional intensity expressed in messages to female agents, such as through the use or avoidance of a ranting style or abusive words.

To test these theoretical predictions and shed further light on how gender affects customer support interactions, we utilized a public dataset of service interactions on Twitter between Southwest Airlines and its customers from March 2018 to September 2019. During this period, customer service agents added their first names to their response messages, which allowed us to infer the gender of these agents. We developed three outcome variables to capture a customer's propensity to further engage after an agent's gender could be inferred and the emotional state of those who continued engaging with agents. To characterize a customer's emotional state, we drew upon the circumplex model of affect in emotion science. Each emotional state is decomposed into the valence dimension and the arousal dimension.

Our identification strategy was based on the Backdoor Criterion (Pearl, 2009) and hinged on two important observations. First, the assignment of customers or, more accurately, customer service requests to agents was largely random because a customer service request was usually assigned to the next available agent without considering the agent gender. This assumption is supported both by anecdotal evidence and the well-balanced customers' characteristics by agent gender. Second, since tweets are short, agents' initial responses were often quite standard, making it feasible to control content variation in those responses, which may have also influenced the outcome variables. Empirical analyses using various content control methods revealed that customers were more likely to continue interacting with female agents than male agents and that their messages were more negative in valence but less intense in arousal. These results are consistent with our theoretical predictions and offer valuable insights both to practitioners and to academics.

2. Related literature

Due to the dominance of female labor in the customer service industry (Steiger & Wardell, 1995), customers' agent gender preferences have been well recognized by academia and industry as playing an important role in offline customer service (Foster & Resnick, 2013). Fischer et al. (1997)'s seminal work suggested two possible causes of customers' preferences for agent gender in offline settings. First, according to gender stereotype bias or gender congruence bias, occupations dominated by a specific gender are perceived to be better suited to people with the characteristics and skills of that gender. Second, according to in-group bias, customers expect agents of the same gender as themselves to provide better service than agents of the opposite gender. The effect of gender congruence bias on customers' evaluation crucially depends on service quality conditions (Luoh & Tsaur, 2007). Under unfavorable service quality conditions, female agents receive lower evaluations from customers than male agents (Hekman et al., 2010; Snipes et al., 2006). By contrast, under favorable service quality conditions, customers tend to perceive higher service quality from female agents than male ones (Luoh & Tsaur, 2007).

Compared to the offline customer service literature, the role of agent gender has thus far not been studied in the online customer service literature. Unlike in offline settings where customers have rich visual and audio cues to infer agent identity, agent identity cues are often limited in online settings, especially in text-based customer service. Very recently, firms started requiring agents to include identity cues (e.g., personal signatures, personal profiles) in their responses to service requests on social media (Cheng & Pan, 2021; Gao et al., 2023). Thanks to this timely new development, the present study closes the research gap by examining the role of agent gender in the delivery of customer service on Twitter.

In recent years, social media customer service has drawn increasing attention from information systems (IS) researchers. One stream of this literature concentrates on the customer perspective, exploring their motivations to complain on platforms like Twitter, preferences between public and private channels, and the effects of agent profiles on complaint willingness (Cheng & Pan, 2021; Gans et al., 2021; He et al., 2023). Outcomes of these service provisions have also been studied, highlighting the importance of identity cues in engagement and resolution (Gao et al., 2023). The second stream of this literature delves into firms' strategies, examining differential responses based on customer data, managing customer sentiment, and linguistic response adjustments (Gunarathne et al., 2018; Mousavi et al., 2020; Proserpio et al., 2021). Our paper belongs to the first stream but with a unique angle

[2]Please note that, unlike Olark's experiment, which was based on private chats, social media customer service is public.

on the role of perceived agent gender by customers, which has not been examined in social media customer service context.

3. Hypothesis development

3.1. Agent gender and customer's willingness to engage

Customer service is an act of helping and emotionally supporting customers. The social psychological literature argued that gender differences in helping behaviors may shape customers' expectations of male and female agents (Bem, 1974; Hoffman, 1977). Women, typically seen as empathetic and compassionate, have historically held the majority of customer service positions (Steiger & Wardell, 1995). Hence, even in the absence of actual service quality by agents of different genders, the dominance of female agents in the customer service industry may have led to the stereotype that females are better at delivering customer service. Indeed, the literature has suggested that an occupation dominated by one gender is stereotyped as being better suited to the characteristics of that gender (Basow, 1992). Therefore, regardless of whether there is any difference in customer services delivered by agents of different genders, customers may prefer female agents due to perceived gender differences in helping behavior or occupational stereotypes.

Hypothesis 1: *After the initial exchange of messages, customers are more likely to continue conversations with female agents than with male agents.*

3.2. Agent gender and customer emotion

In the fast-paced and transparent environment of social media customer service, accurately gauging the multifaceted emotional responses of consumers becomes paramount to both social media managers as well as academics. At the intersection of psychology, communication, and business, we adopted the circumplex (valence-arousal) model as the main metrics for our study. The circumplex model of affect suggests that all emotions arise from two independent neurophysiological systems and can be mapped to points in a two-dimensional space (Posner et al., 2005; Russell, 1980). The valence dimension, visualized in Figure 1 as the horizontal axis, is a pleasure-displeasure dimension. The arousal dimension, visualized in Figure 1 as the vertical axis, is an arousal-sleep or energy dimension. An individual experiencing an emotional state high in the arousal dimension is highly activated and reactive to stimuli. Within this framework, each emotion can be characterized as a distinct point on a circle in this two-dimensional space. For example, the emotional state of feeling serene can be characterized as high valence and low arousal, while the emotional state of feeling furious can be described as low valence and high arousal.

Social media customer service characterized by its fast-pace, immediate, and public interactions necessitates a nuanced measurement of customers' real-time emotional affective shifts. The circumplex model offers a detailed, instantaneous understanding on customers' attitude and potential gender preferences towards agents that broader evaluative measures such as service task resolution often fail to capture. Meanwhile, the inherent transparency of social media platforms deems that agents, in this dynamic environment, are compelled to swiftly and appropriately address customer complaints, often relying on the immediate customer response they perceive. Whereas traditional metrics like customer satisfaction capture post-event cognitive evaluations, the circumplex model emphasizes the immediate, visceral emotional reactions that accurately captures customer opinions and preferences of agents. This model offers a deeper insight into how customers react to agent responses, aiding in the development of adaptive communication strategies that can improve customer experience and bolster brand's digital reputation. Drawing upon this theoretical framework, we analyze how perceived agent gender affects the emotional state of a customer in each dimension. To understand the role of gender in a customer's emotional valence, we first distinguish customers with distinct motivations: goal-oriented and emotion-focused (Kowalski, 1996).

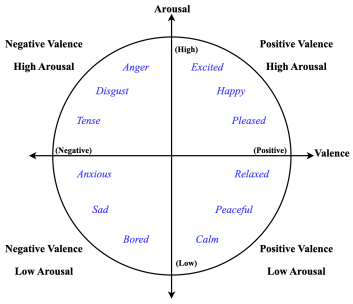

Figure 1. The circumplex model of affects.

For goal-oriented customers, the primary objective is to seek redress or economic compensation through complaining. The emotional valence embedded in their messages serves as an instrument for achieving this objective. These customers may strategically emphasize or exaggerate the negative valence when communicating with female agents because women

are often perceived as more compliant and softhearted than men (Eagly, 1983). Since female agents are perceived as more accommodating than their male counterparts, a goal-oriented customer would optimally exhibit a more negative emotional valence towards a female agent for the best return on investment. For emotion-focused customers, the desire to express emotional dissatisfaction is the primary driver of complaints. These customers may consider female agents to be a better audience because women are often perceived as more empathetic than men (Eagly & Wood, 2012). Emotion-focused customers may expect female agents to better understand and even resonate with their emotions than male agents, making these customers more willing to share their distress with female agents and resulting in more negative valence in their messages. In reality, customers are likely driven to complain by a mix of both motivations. Regardless, the above arguments predict that a customer message addressed to a female agent should exhibit lower emotional valence than one sent to a male agent.

Hypothesis 2: *Conditional on engagement, customers show lower emotional valence towards female agents than male agents.*

To hypothesize the effect of agent gender on a customer's emotional arousal, we propose four potential mechanisms. First, because customers complain to seek redress and/or express dissatisfaction, the perceived ability of agents to understand their feelings is critical to their emotional reaction. Females are often perceived as more capable of understanding others' feelings than males; hence, the presence of a female agent could make a customer feel less agitated and exhibit less emotional arousal. Second, social desirability states that people conform to social norms during social interactions (Chung & Monroe, 2003). Since femininity is traditionally associated with vulnerability (Eagly & Crowley, 1986), females typically receive greater chivalrous treatment than males (FeldmanHall et al., 2016). Therefore, when customers interact with female agents, they are likely to be more respectful and refrained, resulting in a less intense conversation with lower emotional arousal. Third, because women have historically constituted a larger proportion of customer service agents than men (Steiger & Wardell, 1995), the confirmation or disconfirmation of such a belief in the presence of agent gender information may moderate the arousal level of a customer. A customer served by a female agent may feel a sense of ease and familiarity (Vittengl & Holt, 1998), thereby reducing emotional arousal. Finally, unlike the above mechanisms, some customers may exhibit higher emotional arousal when interacting with female agents because women are perceived as more compliant and empathetic (Eagly, 1983), especially when women are emotionally pressured. Although it is unclear which of the above mechanism would dominate in our setting, we predict, simply based on the number of mechanisms, that the overall effect of female gender cues on emotional arousal would be negative.

Hypothesis 3: *Conditional on engagement, customers exhibit lower emotional arousal towards female agents than male agents.*

4. Data and variables

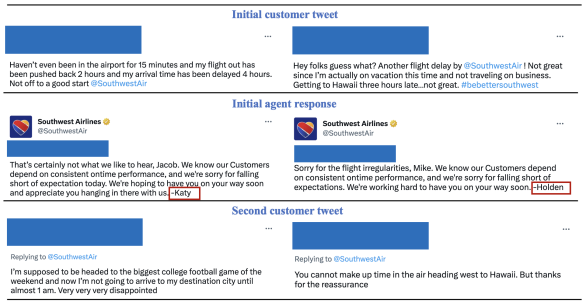

Figure 2. Sample conversations.

We collected all tweets received and posted by Southwest Airlines from March 2018 to September 2019. We leveraged Twitter metadata to reconstruct customer service conversations, starting with initial service requests by customers, followed by back-and-force interactions between agents and customers. We leveraged the support vector machine (SVM) classifier constructed by Gao et al. (2023) to identify customer service-related conversations in the data. Most importantly for our study, an agent signature provides an identity cue through which customers may consciously or subconsciously infer an agent's gender. Figure 2 illustrates two customer service tasks handled by agents with female- and male-dominant names. To infer agent gender from an agent's signature, we hired two annotators to label the dominant gender of each name based on US Census statistics (Word et al., 2008). After excluding conversations with ambiguous or invalid names from the sample, there were 67,736 customer service-related conversations, 23,331 of which continued the conversation following the agent's initial response to the initial service request. Consistent with popular perception, the majority (i.e., 64%) of these agents were female, at least based on gender inferred by names.

Table 1 lists the variables with their definitions and summary statistics. The explanatory variable, *Female*, is

a binary variable equal to one if a customer service task is handled by a female agent. We measure a customer's willingness to engage through a binary variable, *SecondTweet*, which measures whether the customer follows up on the agent's initial response by sending a second tweet. This is a clean, reasonable measure of engagement because the entire conversation usually ends if a customer does not send a second tweet. We then define two outcome variables, *SecondTweetValence* and *SecondTweetArousal*, as continuous measures that capture a customer's emotional state in the second tweet. We only use the second tweet instead of all tweets in the conversation for two reasons. First, the second tweet is the first observable customer reaction after an agent's gender could be inferred from his/her initial response. Therefore, the identification of the gender effect is less vulnerable to potential confounding factors introduced after the second tweet. Second, to measure valence and arousal, we use machine learning algorithms, whose accuracy depends on the text length. Since the number of customer tweets varies from conversation to conversation, using all tweets in a conversation could have led to measurement errors and introduced new confounding factors.

We use two algorithms to construct each outcome variable to alleviate measurement error concerns. For *SecondTweetValence*, we use classifiers based on two supervised machine learning algorithms to generate *SecondTweetValence_LR* and *SecondTweetValence_NB*. Specifically, we randomly selected 5000 customers' messages and hired two annotators to manually label the valence. Based on the labeled data, we chose the logistic regression and the Naïve Bayes classifiers over other classifiers because of their higher performance. For *SecondTweetArousal*, we use supervised learning and lexical-based approaches to construct *SecondTweetArousal_LSTM* and *SecondTweetArousal_Lexical*. Specifically, we built a long short-term memory (LSTM) classifier to predict arousal values based on data provided by Buechel and Hahn (2017) while using the lexicon provided by Warriner et al. (2013) to construct the lexical classifier[3].

5. Identification

To test our hypotheses, we estimate the following econometric model at the conversation level.

$$Y_i = \beta_0 + \beta_1 \text{Female}_i + \beta_2 X_i + \beta_3 Z_t + \epsilon_{it} \quad (1)$$

The outcome variable Y_i is either *SecondTweet*, *SecondTweetValence*, or *SecondTweetArousal* for conversation *i*, depending on which hypothesis is tested. The key coefficient of interest, β_1, captures the main treatment effect of an agent's perceived gender on a customer's engagement or emotional state. X_i includes the characteristics of conversation i, such as the profile of the customer, the content of the first customer tweet, and the content of the first agent response. Z_t includes seasonality variables such as the day of the week and the hour of the day.

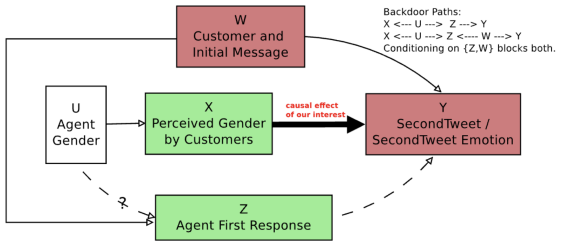

Figure 3. Causal graph and identification strategy.

An ideal research design would involve the random assignment of customer service requests to male or female agents and an identical initial response, other than agent names, from the agents. This would ensure that any statistically significant difference in an outcome variable derived from the second tweet between requests handled by male and female agents is caused by the name-induced gender cue and customers' resulting perceptions. Although such a research design is not possible with observational data, it can be closely approximated thanks to two unique features of our research context. To facilitate the discussion, we illustrate the causal structure among key constructs using the directed acyclic graph (DAG) in Figure 3. In this causal graph, the thick solid line connecting the explanatory variable X (i.e., *Perceived Gender by Customers*) and the outcome variables Y (i.e., *SecondTweet*, *SecondTweetValence*, and *SecondTweetArousal*) represents the main causal effect of interest (i.e., $X \rightarrow Y$ in the causal graph).

We first argue that there is no arc connecting node W and node U because, in large organizations such as Southwest Airlines, a customer complaint is usually assigned to the next available agent, which should be largely independent of the agent's gender, especially after controlling for seasonality (e.g., the day of the week, the hour of the day). This makes sense for two reasons. First, customers care about the speed of response and firms care about workforce efficiency, which makes it beneficial to assign a service request to the next available agent, regardless of the agent's gender. Indeed, the Next Available Agent (NAA) strategy is commonly used by call centers to reduce customer wait

[3]Due to the occurrence of terms within the emotional lexicon, the lexical classifier constructed arousal weights for 65,574 out of 67,736 tweets

Table 1. Summary statistics.

Variable	Mean	S. D.	Definition
Treatment			
Female	0.74	0.44	Binary variable indicating whether an agent is female
Outcome Measures			
SecondTweet	0.33	0.47	Binary variable indicating whether a customer engages with an agent's first reply
SecondTweetValence_LR	0.67	0.20	Continuous measure of customers' emotional valence in the second tweet based on the Logistic Regression classifier (within [0,1])
SecondTweetValence_NB	0.65	0.19	Continuous measure of customers' emotional valence in the second tweet based on the Naïve Bayes classifier (within [0,1])
SecondTweetArousal_LSTM	0.71	0.01	Continuous measure of customers' emotional arousal in the second tweet based on the LSTM classifier (within [0,1])
SecondTweetArousal_Lexical	0.66	0.10	Continuous measure of customers' emotional arousal in the second tweet based on lexical weights (within [0,1])
Customer Characteristics			
LogFollowers	5.07	1.97	Log-transformed number of followers a customer has
LogFollowings	5.62	1.47	Log-transformed number of followings of a customer
LogUpdates	7.30	2.44	Log-transformed number of tweets of a customer
InitialValence_LR	0.58	0.16	Continuous measure of customers' emotional valence in the initial tweet based on the Logistic Regression classifier (within [0,1])
InitialValence_NB	0.57	0.15	Continuous measure of customers' emotional valence in the initial tweet based on the Naïve Bayes classifier (within [0,1])
InitialArousal_LSTM	0.72	0.01	Continuous measure of customers' emotional valence in the initial tweet based on the LSTM classifier (within [0,1])
InitialArousal_Lexical	0.68	0.11	Continuous measure of customers' emotional valence in the initial tweet based on the lexical weights (within [0,1])
InitialAggression	0.10	0.12	Continuous measure of a customer's aggression in the initial tweet
InitialWords	26.40	13.04	Number of words in a customer's initial Tweet
InitialHello	0.11	0.31	Binary variable indicating whether a customer greets an agent
InitialGratitude	0.07	0.27	Binary variable indicating whether a customer is polite in their initial request using words such as "thank you", "I appreciate", etc.
InitialBareCommand	0.19	0.41	Binary variable indicating whether a customer expresses the service request in grammatically polite structure (i.e., starts a sentence with the unconjugated verbs)
InitialQuestions	0.16	0.40	Binary variable indicating the existence of question marks in the initiated tweet
Agent Reply Quality			
AgentWords	28.98	13.88	Number of words in an agent's first reply
ResponseTime	0.43	1.01	Number of hours it takes for an agent to respond
DM	0.29	0.45	Binary variable indicating whether an agent mentions direct messages
Please	0.28	0.45	Binary variable indicating whether an agent mentions "please"
Hello	0.33	0.47	Binary variable indicating whether an agent greets a customer
Apology	0.47	0.50	Binary variable indicating whether an agent apologizes to a customer
Hedges	0.36	0.48	Binary variable indicating whether an agent shows uncertainty

Note. This table reports the summary statistics of the key variables. S.D. stands for standard deviation. The number of observations is 67,736. InitialValence_LR is controlled for when SecondTweetValence_LR is the outcome variable, and InitialValence_NB is controlled for when SecondTweetValence_NB is the outcome variable. Similarly, InitialArousal_LSTM is included when SecondTweetArousal_LSTM is the outcome variable, InitialArousal_Lexical is included when SecondTweetArousal_Lexical is the outcome variable.

time and optimize agent productivity. Second, the practice of assigning service requests to agents based on their gender is at best controversial, and at worst legally problematic. Therefore, we can think of the assignment of customers to the two groups of agents (i.e., male and female) as independent of the group ID, much like in a randomized experiment. To empirically evaluate the plausibility of this assumption, we check the balance of customer characteristics and the content of the customer's initial tweet across the two groups. We present the summary statistics and definitions of customer characteristics in Table 1. We find that the standardized differences are all below the threshold of 0.10 suggested by Austin (2009) to assess covariate balance for samples generated from matching or from randomized experiments. Therefore, we believe that the assignment process is largely random and that any unobserved customer characteristic is likely balanced across the two groups.

Recall blocking means stopping the flow of dependency between nodes connected by paths. Formally, given three disjoint subsets of nodes X, Y, and Z, a path p connecting nodes in X and Y is said to be **blocked** by Z if and only p contains a **chain**

($i \rightarrow z \rightarrow j$) or a **fork** ($i \leftarrow z \rightarrow j$) such that $z \in Z$, or if p contains a **collider** ($i \rightarrow m \leftarrow j$) such that $m \notin Z$ and no descendants of m is in Z.

In our causal graph, there are two potential backdoor paths, where a backdoor path between a causal node and an outcome node is defined as any path with an arrow into the causal node. The first backdoor path (i.e., *Perceived Gender by Customers* $\leftarrow$ *Agent Gender* $\rightarrow$ *Agent First Response* $\rightarrow$ *Second Tweet / Second Tweet Emotion*), connected through the dashed line at the bottom of Figure 3, may create a spurious correlation between perceived agent gender and the outcome variables through the mechanism of different initial responses by agents of different genders. To block this backdoor path, we need to control for an agent's initial response. While it is straightforward to include the response time (i.e., the elapsed time between a customer's initial tweet and an agent's initial response) as a control, it is less clear how to control for the content of an agent's initial response due to the complexity of natural language. We believe this is achievable thanks to the second feature of our research context: the lack of significant, meaningful variation in the first response by customer service agents. Moreover, the short text length of the first response allows us to control for its content sufficiently well so that the potential bias caused by any residue is likely small. We use several approaches to control the content of an agent's first response. The first is the traditional approach of feature extraction in which we include features explicitly constructed to characterize agent response. We present the summary statistics and definitions of agent response characteristics in Table 1. Our second approach involves creating an embedding vector for each agent response. Specifically, we employ latent semantic analysis (LSA) to create a 20-dimensional dense vector for each initial response in which each element of the vector is a continuous, rather than discrete, measure of the response in some latent semantic dimension. The idea of embedding has been widely used in natural language processing over the past two decades, and its value for causal inference has been recognized in recent years. In our third approach, we employ clustering to categorize content types, which have been recently exploited for identification. Specifically, we generate the document-term matrix for all agents' responses and perform singular value decomposition (SVD) to create word embedding. Next, we apply k-means clustering on word embedding to create agent reply clusters. We use the silhouette score to determine the optimal number of clusters, with 15 clusters generating the highest silhouette score. Finally, we include cluster dummies in the model as the control for content.

The second potential backdoor path (i.e., *Perceived Gender by Customers* $\leftarrow$ *Agent Gender* $\rightarrow$ *Agent First Response* $\leftarrow$ *Customer and Initial Complaint* $\rightarrow$ *Second Tweet / Second Tweet Emotion*) is blocked if we control for *Customer and Initial Complaint*, which is a fork. Note that *Agent First Response* is a collider here. Since we want to control for this node to block the first potential backdoor path, we have to block the second potential backdoor path by controlling for the fork node. Similar to our control for *Agent First Response*, we can use various strategies to control for characteristics of the customer and the initial complaint. Note that if agent gender only affects agent first response in a way that does not affect the outcome node (i.e., the backdoor path from U to Y through Z is broken), then we could identify the causal effect simply by not controlling for *Agent First Response* because it is a collider in the second potential backdoor path. In fact, the results of this simplified identification strategy are precisely what the model-free comparison shows[4].

In summary, if we condition on the nodes *Agent First Response* and *Customer and Initial Complaint*, both potential backdoor paths are blocked. Finally, recall the Backdoor Criterion (Pearl, 2009) that for three disjoint sets of nodes, $\mathcal{D}$, $\mathcal{Y}$, and X, the causal effect of $\mathcal{D}$ on $\mathcal{Y}$ can be identified if, for any $D \in \mathcal{D}$, $Y \in \mathcal{Y}$, the set of controls X contains no descendant of D and blocks every backdoor path between D and Y. Therefore, we can use observable data to identify our causal effects of interest.

6. Main results

Since we may improve statistical inference with a matched sample in the absence of a completely balanced sample, we perform propensity score matching (PSM) with one-to-one matching using customer and agent characteristics before the second tweet as matching variables and a caliper of 0.001 following Austin (2011). Table 2 reports the estimation results with *SecondTweet* as the outcome variable. The differences among the first three columns reflect the three strategies (i.e., feature-based control, embedding, and cluster dummies) used to control the content of an agent's first response. Despite the differences in sampling and content control strategies, the estimated coefficient of the *Female* variable is significantly negative ($p < 0.01$). Hence, customers are more likely to continue a service conversation with female agents than with male agents, thereby supporting **Hypothesis 1**.

Table 3 reports the estimation results with

[4]The results of the model-free comparison are available upon request.

Table 2. Agent gender on customer engagement.

	SecondTweet			
	Full	Full	Full	PSM
	(1)	(2)	(3)	(4)
Female	0.0393***	0.0364***	0.0285***	0.0410***
	(0.0040)	(0.0042)	(0.0040)	(0.0043)
Agent Controls	Y			Y
Word Embedding		Y		
Response Cluster FE			Y	
Observations	67736	67736	67736	63993
R^2	0.0424	0.0500	0.0411	0.0425

Note. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are reported in parentheses. Customer controls, agent response time, and seasonality FE are included in all specifications.

Table 3. Agent gender on customer second tweet valence.

	SecondTweetValence_LR				*SecondTweetValence_NB*			
	Full	Full	Full	PSM	Full	Full	Full	PSM
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-0.0121***	-0.0119***	-0.0114***	-0.0130***	-0.0136***	-0.0131***	-0.0123***	-0.0144***
	(0.0029)	(0.0029)	(0.0029)	(0.0029)	(0.0028)	(0.0028)	(0.0028)	(0.0028)
Agent Controls	Y			Y	Y			Y
Word Embedding		Y				Y		
Response Cluster FE			Y				Y	
Observations	22331	22331	22331	21308	22331	22331	22331	21308
R^2	0.1445	0.1953	0.1455	0.1466	0.1451	0.1902	0.1450	0.1465

Note. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are reported in parentheses. Customer controls, agent response time, and seasonality FE are included in all specifications.

SecondTweetValence as the outcome variable, which is operationalized using two algorithms. The first four columns report the estimation results when the second tweet valence is measured by *SecondTweetValence_LR*, while the last four columns report the estimation results when the valence is measured by *SecondTweetValence_NB*. The last column of each group (i.e., Columns 4 and 8) corresponds to the matched sample, while the other three correspond to the full sample with different strategies for content control of an agent's first response. As evident in the table, regardless of the differences in measure construction, sampling strategy, and content control, the estimated coefficients of *Female* are significantly negative ($p < 0.01$), indicating that customers' emotional valence is lower when their service requests are handled by a female agent than by a male agent. Therefore, **Hypothesis 2** is supported.

Table 4 reports the estimation results with *SecondTweetArousal* as the outcome variable. Similar to Table 3, Table 4 contains two constructions of the outcome variables, two samples, and three strategies to control for the content of an agent's first response. The negative estimated coefficients ($p < 0.01$) of the *Female* variable across all specifications suggest that a customer's emotional arousal is reduced in the presence of a female agent, thereby supporting **Hypothesis 3**.

7. Conclusions

This paper studies the effect of perceived agent gender on customer behavior in the context of social media customer service. Using a unique data set with rich conversation details, we found that customers are more willing to engage with female agents than male agents in customer service conversations. Based on the theoretical decomposition of emotion into valence and arousal, we found empirical evidence that, upon engagement, customers show lower emotional valence and arousal towards female agents than male agents.

This work is the first academic study to empirically examine how the perceived gender of customer service agents affects customer behavior in an online environment while it is also among the first in the IS field to evaluate customer emotion along the valence and arousal dimensions in online customer service, especially by leveraging machine learning techniques. Even though the circumplex model of affect (Posner et al., 2005; Russell, 1980) suggests that arousal is as critical as valence, the former is much less studied by researchers, partly due to the lack of methodologies to accurately measure this dimension. The significant effects of perceived agent gender on these indispensable dimensions in our paper echo the theoretical framework and corroborate the importance of incorporating arousal when analyzing customer emotion.

Table 4. Agent gender on customer second tweet arousal

	SecondTweetArousal_LSTM				*SecondTweetArousal_Lexical*			
	Full	Full	Full	PSM	Full	Full	Full	PSM
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-0.0006***	-0.0006***	-0.0005***	-0.0005***	-0.0024**	-0.0019*	-0.0020**	-0.0028**
	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0010)	(0.0011)	(0.0010)	(0.0011)
Agent Controls	Y			Y	Y			Y
Word Embedding		Y				Y		
Response Cluster FE			Y				Y	
Observations	22331	22331	22331	21308	20280	20280	20280	19348
R^2	0.0099	0.0123	0.0082	0.0101	0.0108	0.0196	0.0130	0.0110

Note. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors are reported in parentheses. Customer controls, agent response time, and seasonality FE are included in all specifications.

Our paper also provides valuable insights to practitioners, especially to managers who strive to improve the equity and equality of the service industry workforce. Our findings demonstrate that customers' preferences concerning agent gender indeed exist and their differential behaviors are real. As a result, it is imperative to take action to alleviate the potential negative consequences of gender inequality. In particular, female agents may experience a higher workload and a greater emotional burden, which may lead to inferior service performance evaluation and ultimately affect their chances for promotion and bonuses. In addition to customers' emotional quantiles, we also conducted an analysis on customers' satisfaction at the end of the conversation. We found consistent results that customers are less likely to express appreciation (or feel satisfied) towards female agents compared to male agents, conditional on customer characteristics and agents' service quality. These results are available upon request. Hence, female agents' incentives and job satisfaction may decline over time, leading to higher attrition rates. Therefore, managers should consider customers' gender preferences and the toll on female agents when evaluating service performance. For instance, common assessment metrics, such as the number of requests handled, average response time, and customer sentiment, may underestimate female agents' service effectiveness. Instead, performance-based assessments should additionally account for the emotional overload imposed on female agents by customers. Alternatively, providing emotional support and training to female agents could help retain the best talent and, more importantly, create a more equitable and inclusive working environment.

Considering the increasing popularity of using chatbot to deliver customer service, especially after the release of ChatGPT and GPT-4, exploring the role of perceived agent "gender" could reveal just how deep-rooted our gender preference goes. Such an understanding can help with the design of anthropomorphized AI agents in customer service, complementing recent findings (e.g., Han et al. (2022)) on chatbots with human emotions.

References

Austin, P. C. (2009). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics-simulation and Computation*, *38*(6), 1228–1234.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*(2), 150–161.

Basow, S. A. (1992). *Gender: Stereotypes and roles*. Thomson Brooks/Cole Publishing Co, Belmont, CA.

Bem, S. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42*(2), 155–162.

Buechel, S., & Hahn, U. (2017). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 578–585. https : / / aclanthology . org / E17-2092

Cheng, H.-T., & Pan, Y. (2021). I'm not a chatbot: An empirical investigation of humanized profiles of social media customer service representatives. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 4167–4176.

Chung, J., & Monroe, G. S. (2003). Exploring social desirability bias. *Journal of Business Ethics*, *44*(4), 291–302.

Eagly, A. H. (1983). Gender and social influence: A social psychological analysis. *American Psychologist*, *38*(9), 971–981.

Eagly, A. H., & Crowley, M. (1986). Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin*, *100*(3), 283–308.

Eagly, A. H., & Wood, W. (2012). Social role theory. Van Lange PA, Higgins ET, Kruglanski AW, eds. *Handbook of Theories in Social Psychology*, *2*, 458–476.

FeldmanHall, O., Dalgleish, T., Evans, D., Navrady, L., Tedeschi, E., & Mobbs, D. (2016). Moral chivalry: Gender and harm sensitivity predict costly altruism. *Social Psychological and Personality Science*, *7*(6), 542–551.

Fischer, E., Gainer, B., & Bristor, J. (1997). The sex of the service provider: Does it influence perceptions of service quality? *Journal of Retailing*, *73*(3), 361–382.

Foster, C., & Resnick, S. (2013). Service worker appearance and the retail service encounter: The influence of gender and age. *The Service Industries Journal*, *33*(2), 236–247.

Gans, J. S., Goldfarb, A., & Lederman, M. (2021). Exit, tweets, and loyalty. *American Economic Journal: Microeconomics*, *13*(2), 68–112.

Gao, Y., Rui, H., & Sun, S. (2023). The power of identity cues in text-based customer service: Evidence from Twitter. *MIS Quarterly*, *Forthcoming*.

Gunarathne, P., Rui, H., & Seidmann, A. (2018). When social media delivers customer service: Differential customer treatment in the airline industry. *MIS Quarterly*, *42*(2), 489–520.

Han, E., Yin, D., & Zhang, H. (2022). Bots with feelings: Should AI agents express positive emotion in customer service? *Information Systems Research*, *Forthcoming*.

He, S., Lee, S.-Y., & Rui, H. (2023). Open voice or private message? The hidden tug-of-war on social media customer service. *Production and Operations Management*, *Forthcoming*.

Hekman, D. R., Aquino, K., Owens, B. P., Mitchell, T. R., Schilpzand, P., & Leavitt, K. (2010). An examination of whether and how racial and gender biases influence customer satisfaction. *Academy of Management Journal*, *53*(2), 238–264.

Hoffman, M. (1977). Sex differences in empathy and related behaviors. *Psychological Bulletin*, *84*(4), 712–722.

Kowalski, R. M. (1996). Complaints and complaining: Functions, antecedents, and consequences. *Psychological Bulletin*, *119*(2), 179–196.

Luoh, H.-F., & Tsaur, S.-H. (2007). Gender stereotypes and service quality in customer–waitperson encounters. *Total Quality Management*, *18*(9), 1035–1054.

Mousavi, R., Johar, M., & Mookerjee, V. S. (2020). The voice of the customer: Managing customer care in Twitter. *Information Systems Research*, *31*(2), 340–360.

Pearl, J. (2009). *Causality: Models, reasoning and inference, 2nd ed.* Cambridge University Press. New York, NY, USA.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734.

Proserpio, D., Troncoso, I., & Valsesia, F. (2021). Does gender matter? The effect of management responses on reviewing behavior. *Marketing Science*, *40*(6), 1199–1213.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Snipes, R. L., Thomson, N. F., & Oswald, S. L. (2006). Gender bias in customer evaluations of service quality: An empirical investigation. *Journal of Services Marketing*, *20*(4), 274–284.

Steiger, T. L., & Wardell, M. (1995). Gender and employment in the service sector. *Social Problems*, *42*(1), 91–123.

Vittengl, J. R., & Holt, C. S. (1998). Positive and negative affect in social interactions as a function of partner familiarity, quality of communication, and social anxiety. *Journal of Social and Clinical Psychology*, *17*(2), 196–208.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, *45*, 1191–1207.

Word, D. L., Coleman, C. D., Nunziata, R., & Kominski, R. (2008). Demographic aspects of surnames from census 2000.