Online Review Censorship

Aida sanatizadeh
University of Illinois at Chicago, asanat2@uic.edu

Gordon Burtch
Boston University, gburtch@bu.edu

Yili Hong
University of Houston, yhong@central.uh.edu

Yuheng Hu
University of Illinois at Chicago, henryyhu@gmail.com

Online Review Censorship

Completed Research Paper

Aida Sanatizadeh
University of Illinois Chicago
Chicago, IL 60607
asanat2@uic.edu

Gordon Burtch
Boston University
Boston, MA 02215
gburtch@bu.edu

Yili Hong
University of Miami
Coral Gables, FL 33146
khong@miami.edu

Yuheng Hu
University of Illinois Chicago
Chicago, IL 60607
yuhenghu@uic.edu

Abstract

Ample anecdotal evidence in the media notes that many businesses seek to 'silence' negative reviews, e.g., via legal threat. Despite attention toward this issue, we are aware of no systematic analyses addressing it. We address that gap here, leveraging review data from TripAdvisor.com. First, we estimate that ~1% of truthful reviews are deleted within six months of posting and that negative reviews are significantly more likely to be deleted, consistent with a mechanism of censorship. The effect is substantial; we estimate that a 1-star decrease in rating valence is associated with an approximate 25% (0.25 pp) increase in the probability of deletion. Second, we examine how freedom of expression (FoE) in a country associates with characteristics of (uncensored) online reviews. We find that FoE associates with larger review volumes, lower review valence, and faster review posting. We discuss implications for online ratings platforms, consumers, and research opportunities

Keywords: Online reviews, censorship, self-censorship

Introduction

Online reviews are of critical importance to retailers. Prior studies and industry reports indicate that online ratings drive an average 18% increase in sales (Luca 2016) and that 86% of consumers would hesitate to purchase from a business that has negative online reviews (Unni 2020). Given the importance of online reviews, it is not surprising that many businesses go to great lengths to obtain and maintain positive ratings. Numerous media reports in recent years speak to businesses' efforts to censor negative consumer opinions, either by offering consumers an incentive or refund to remove a bad review or by making threats, e.g., of a lawsuit. Examples are the case in New York where a gynecologist filed a $1 million defamation lawsuit against a patient over a negative online review (Moutos et al. 2020) and the case when a restaurant owner threatened to sue the customer for defamation unless the customer removes negative reviews from Yelp.[1] In response to this issue, review platforms and regulators have taken note and sought to curb the problem. Yelp was one of the pioneers that launched a program that aimed to detect businesses' abnormal activities.

Review censorship has undesirable consequences that parallel those of fake reviews (Luca and Zervas 2016), ultimately misleading consumers about other consumers' experiences with a retailer or service provider. Moreover, censorship eliminates those instances of feedback that consumers are known to rely on most

[1] https://thisisreno.com/2019/09/negative-online-reviews-and-legal-threats-opinion

heavily; according to previous literature, consumers place greater weight on negative reviews (Mayzlin et al. 2014; Sen and Lerman 2007; Yin et al. 2014). Review platform operators thus also have a strong incentive to prevent review censorship, as it undermines consumer trust in the platform's content.

In 2020, Yelp introduced a Questionable Legal Threats Alert, flagging the pages of businesses that attempt to pressure consumers to remove negative reviews via the threat of legal action. In response to this issue, the Consumer Financial Protection Bureau (CFPB) issued a policy statement, indicating that "in America, no corporation should be able to silence a customer from posting an honest review online." [2] The CFPB added that "corporate disinformation campaigns that suppress legitimate reviews… are not only a threat to free speech and fair competition, but they are also illegal." Nonetheless, the issue persists globally, as the media continues to report instances of consumers being threatened and even jailed for authoring negative reviews. As one famous, recent example, in 2020, the New York Times reported on an American guest at a Thai hotel who was arrested and jailed after refusing to delete a negative review they had posted to TripAdvisor.[3]

Despite the attention this issue has received, large-scale investigations are currently lacking around the prevalence of review censorship, either in general or across contexts. Since this information is crucial to informing any platform or policy response, we thus seek to address the gap here. More formally, we address the following two research questions:

RQ1: How prevalent is online review censorship, and how does the prevalence of review censorship vary with the characteristics of an online review?

RQ2: How does the local social climate around freedom of speech relate to the characteristics and volume of online reviews that consumers produce?

We empirically evaluate these questions leveraging a large-scale dataset of hotel reviews collected from TripAdvisor.com. We first observe that approximately 1% of all truthful reviews are deleted within 6 months of posting; an approximate measure of the prevalence of review censorship. Subsequently, our regressions demonstrate that review deletion is most likely to occur when rating valence is low and textual sentiment is negative, as we would expect. Our results demonstrate that review censorship is associated negatively with the rating valence (β = -0.0025, p-value <0.01) and the review sentiment (β = -0.005, p-value <0.01) received by a hotel. Second, we consider the broader relationships between freedom of speech in a country and aggregate online rating distributions. We show that reduced freedom of speech 1) associates negatively with the volume or number of reviews (β = -0.357, p-value <0.01) received by a hotel, 2) associates positively with the rating valence (β_1=0.00132, p-value<0.05), and review sentiment (β_1=0.000289, p-value <0.1) received by a hotel, and 3) associated with lengthier delays in the timing of review posting (β_1=0.00146, p-value <0.01), presumably as consumers wait to depart the hotel's jurisdiction before reporting their feedback.

Our work contributes to the prior literature in several ways. First, we demonstrate that almost 1% of truthful reviews were deleted after being posted presumably due to business owners' threats to consumers over their negative reviews.[4] No prior work has quantified the prevalence of censorship on truthful reviews. Second, we show how review censorship is associated with review characteristics. Previous studies mostly show the association between fake reviews with review characteristics (He et al. 2022). Third, drawing on the influential factors' impact on online reviews, we formulate and evaluate several hypotheses using a unique dataset that integrates online hotel reviews from TripAdvisor with country-level measures of free speech values. We examine the relationship between freedom of speech indices with the review features including rating valence, sentiment, volume, and the delay in posting the consumers' reviews. Previous studies have mostly relied on the role of cultural differences in determining the volume and rating valence of reviews (Fang et al. 2013; Hong et al. 2016).

Our work also has important managerial implications. We show that review deletion is most likely to occur when rating valence is low and textual sentiment is negative, which may misrepresent a hotel's performance and mislead consumers in their purchasing. In response to this issue, platforms can effectively embed a built-in penalty for businesses that make a threat to their consumers by making this fraudulent activity information available to consumers. Recent studies suggest that 80% of users trust a review platform more

[2] https://consumerfinance.gov/about-us/newsroom/cfpb-issues-policy-on-contractual-gag-clauses-and-fake-review-fraud

[3] https://nytimes.com/2020/11/11/world/asia/thailand-hotel-tripadvisor-jail.html

[4] https://tripadvisor.com/ShowTopic-g1-i12105-k7503052-Review_threats_from_owner-Tripadvisor_Support.html

if it displays fake reviews (Ananthakrishnan et al. 2020). This can benefit businesses as well. A platform may admit to users that a business has committed fraud, but that is balanced by an increase in trust from consumers who now see that the business is doing something to address that issue. Businesses should pay attention that customers' reviews are an important way to promote services in a competitive market and if reviews become unreliable, businesses may lose customers which can directly affect sales turnover. Further, our findings show that countries' freedom of expression can be considered a country-specific characteristic that affects different dimensions of reviews. Platforms can provide freedom of speech scores of countries where the businesses are located on their website for users.

The remainder of the paper is organized as follows. We first review related work dealing with online reviews, particularly work addressing review manipulation, motivating several hypotheses. Second, we present our data sources and measures. Third, we introduce our research design and empirical specification to test our hypotheses. We then present our results. Finally, we discuss our findings, and their implications, before concluding with an overview of opportunities for future research.

Literature Review

Online reviews now play a central role in consumer decision-making (Aral 2014). Reviews provide pre-purchase information on almost all products and reveal crowd opinions on those in categories neglected by professional critics (Reimers and Waldfogel 2021). Reviews enable consumers to share their opinions of products or services with other potential buyers (Lee and Bradlow 2011; Lu et al. 2013; Dellarocas 2003). Further, online reviews can help businesses increase their sales significantly (Chevalier and Mayzlin 2006).

Strongly positive or negative reviews are more helpful for search goods than moderate reviews (Mudambi and Schuff 2010). In particular, Park and Lee (2009), and Yin et al. (2014) demonstrate that negative emotions of reviews are more useful for users than other emotions in purchasing a product (Park and Lee 2009; Yin et al. 2014). However, many businesses do not stand negative reviews. Several significant litigation cases have arisen over negative reviews in recent years, where business owners have taken consumers to court over claims of defamation.

Unsurprisingly, businesses seek to manipulate online reviews, and lots of work has now documented such manipulation attempts as they relate to buying fake reviews. Businesses are known to 'silence' or 'censor' consumer opinions expressed in reviews (Mayzlin et al. 2014). A wide array of products purchase fake reviews that include many reviews with an extremely positive ratings (He et al. 2022). Fake reviews are a type of review manipulation that can mislead and deceive communication in a digital environment. They can damage the credibility of reviews and negatively affect businesses' reputations (Luca and Zervas 2016; Lappas et al. 2016). While several prior works have examined the question of paid, fake reviews, no prior work has considered the prevalence and businesses' attempts to censor truthful reviews.

Stevens et al. (2021) demonstrated that censoring negative online reviews can weaken the consumer's brand perceptions (Stevens et al. 2021). Brands need to have a purposeful response strategy in place when dealing with negative sentiments online (Stevens et al. 2018). For example, inserting an empathy statement in response to a negative review or including a paraphrase of the complaint to make the response more personal and less generic, can improve the ratings of the response (Min et al. 2015; Allard et al. 2020). Management responses can mitigate the adverse effects of negative reviews with an apology with assurance or an apology with correction action (Rose and Blodgett 2016; Danescu-Niculescu-Mizil et al. 2009). However, some businesses attempt to control their consumers' online reviews by instituting "gag clauses" in consumers' contracts (Ponte 2016). They intentionally place provisions buried deep in the consumers' contracts to silence potential critics through legal threats and to tip the scales in their favor when it comes to building their online reputations. This stands in direct conflict with the review platforms such as Yelp and TripAdvisor which provide consumers access to reliable and useful information. Restrictive clauses not only impact one's right to free speech but also impact others not to see honest views (Ponte 2016). Further, businesses that employ "gag clauses" may unfairly achieve high overall ratings and appear to do better in the eyes of consumers as opposed to their competitors who do not use them and may receive mixed reviews on crowdsourced review websites. According to CFPB, companies that include clauses in form contracts that forbid a consumer from posting an honest review, limit free speech and therefore are engaged in unfair or deceptive practices.[2] In response to this problem, in 2016, the federal Consumer Review Fairness Act issued a new law that prohibits manipulating reviews including 1) limiting the posting of negative reviews, and 2) imposing contractual ‘gag’ clauses on consumers’ forms of contracts that prevent honest reviews (Calvert 2018). CFPB seeks to ensure that customers can write reviews, particularly ones posted online, about companies’ products and services that accurately reflect their opinions and experiences.

Hypothesis Development

We propose several research hypotheses for empirical testing. We divide the research framework into two groups 1) review censorship and 2) freedom of expression. For the first group, review deletion is the dependent variable (DV), and we empirically examine how review characteristics are associated with the review deletion. For the second group, the freedom of expression value is the independent variable and we examine how it associates with the review characteristics.

Review Censorship

It has been observed that businesses frequently seek to manipulate consumer quality perceptions through the purchase of fake reviews (Mayzlin et al. 2014; Luca and Zervas 2016). And, unfortunately, recent work also documents that these manipulation efforts are effective; after engaging in review buying, once retailers cease, their average ratings fall and the number of one-star reviews they receive begins to increase significantly (He et al. 2022). Our work addresses the complementary problem of review censorship, wherein businesses may actively seek to remove truthful negative feedback. As prior work has documented evidence of paid positive reviews for a business, as well as paid negative reviews for competitors (Luca and Zervas 2016), it is reasonable to expect that review censorship will follow similar patterns. In particular, we expect that review censorship will be most prevalent for low-rating valence reviews and reviews exhibiting negative textual sentiment. Thus, we propose our first two hypotheses:

H1: Reviews are more likely to be deleted when they are of low rating valence.

H2: Reviews are more likely to be deleted when they exhibit negative textual sentiment.

Freedom of Expression

Freedom of expression refers to a person’s right to articulate their views and opinions without fear of retaliation, censorship, or legal sanction. The climate related to freedom of expression varies widely across countries. While some of this variation is due to culture, a great deal is also related to the culture around defamation and litigiousness. Corporations have long been known to employ strategic lawsuits against public participation, or SLAPP lawsuits, to silence their detractors, i.e., to censor consumers. The extent to which such censorship occurs varies across locations, as a function of the local legal context, e.g., whether defamation is criminalized in a particular jurisdiction, and the ease with which such lawsuits can be filed and pursued.

Several prior studies in the context of online reviews have observed that consumers from different countries rate the same product differently, even on the same platform, presumably due to differences in cultural background and norms (Danescu-Niculescu-Mizil et al. 2009; Nakayama and Wan 2019). For example, consumers from collectivist cultures, as in China, are less likely to write reviews with low rating valence such as 1-star ratings (Fang et al. 2013). Here, we consider the potential that such variation may be due to more than mere cultural differences, but instead to differences in the prevalence of censorship attempts.

Online censorship has been studied previously in the context of social network sites, particularly in the form of self-censorship. In social networking contexts, users must make choices about whether and when to share content. Prior has documented that the social environment may lead users to self-censor, presumably out of a desire to avoid negative social repercussions (Sleeper et al. 2013). Individuals who reside in countries that lack freedom of speech protections may nonetheless be worried about the consequences of even posting an anonymous review. According to Sleeper et. Al (2013), people most likely self-censor content followed by personal opinions and updates and are less likely self-censor if they were able to target their audiences (Sleeper et al. 2013). They may fear, for example, that their identity may be discovered. As such, while the effects may be less pronounced than what has been seen in other, identity-verified settings, the threat exists nonetheless and is expected to impede review production.

Thus, consumers may either self-censor their feedback or remove posted feedback in the face of pressure from rated retailers. Given that some jurisdictions lend themselves more toward concerns for legal or social repercussions from posting negative feedback, we expect consumers from countries with greater restrictions on freedom of expression will tend to withhold their negative ratings, and thus report systematically higher ratings on average. Similarly, we expect they will tend to express more positive sentiments in their reviews. Thus, we propose the following two hypotheses:

H3: Freedom of speech restrictions (reduced freedom of speech) will lead to fewer negative reviews, and thus increased average rating valence of (visible, uncensored) reviews.

H4: Freedom of speech restrictions will lead to more positive reviews, and thus increased textual sentiment of (visible, uncensored) reviews.

By the same logic, we might also expect that in countries where restrictions on freedom of expression are higher, consumers will be more conservative in the opinions they post online. In particular, we consider that consumers may be more likely to post feedback only after they depart a hotel or country, to avoid negative repercussions. We thus propose our fifth hypothesis, as follows:

H5: Freedom of speech restrictions will cause more delays in posting reviews, and thus increase the time of posting reviews.

Finally, censorship is, ultimately, the suppression of speech. We thus might expect, as an extension of the above notion of posting delays, that review posting should be less likely to occur, in general, as negative feedback is eliminated. Accordingly, we expect that hotels located in countries with higher restrictions on freedom of expression will tend to receive fewer reviews. This leads us to our final hypothesis:

H6: Restrictions on freedom of expression in a hotel's location will result in fewer reviews being authored, and thus a lower total volume of reviews.

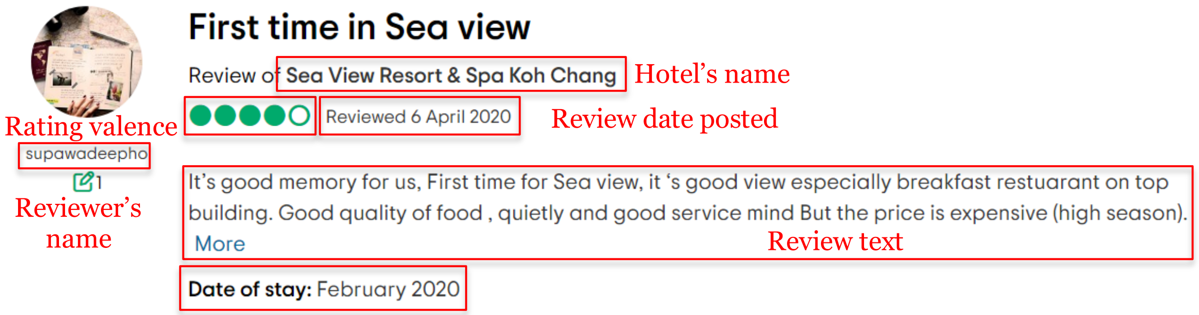
Methodology

Data Collection

In June of 2021, we collected a large sample of online reviews posted to TripAdvisor. We focused our attention on 225 major global hotel chains in this data collection effort, obtaining all reviews dating back to 2013, on every hotel location associated with each chain. We chose 170 major worldwide hotel chains because they maintain a high degree of standardization in service delivery as part of their franchising agreements, despite being independently owned and operated. Accordingly, when it comes to our aggregate analyses, we can exploit variation in rating distributions within a hotel brand, across geographies to isolate variation arising from freedom of speech differences across locations.

For each hotel, we collect the hotel name/chain, quality level (e.g., 3-star hotel), address, textual description, and historical review volume. For each associated review, we collect the review's URL, rating valence (an integer between 1 and 5), review text, the author's profile name, the date the review was posted, and the reported date of the stay (see Figure 1). Using the review text, we construct a measure of sentiment. To compute the sentiment, we employed the VADER (Valence Aware Dictionary and sEntiment Reasoner) library, a lexicon and rule-based sentiment analysis tool that is specifically attuned to a sentiment expressed in social media. To ensure we can calculate this sentiment measure, we focus our analyses on English-language reviews. The final dataset includes more than 7 million reviews.

Next, for each review that was originally posted between January and June of 2021, we revisited its URL in late 2021. If a review was no longer present on TripAdvisor.com, we marked it as deleted. Among 115,487 reviews that were posted between January and June of 2021, approximately 1% were deleted. In addition to the TripAdvisor review data, we also collected annual data on freedom of expression in 180 countries maintained by Reporters Without Borders (www.rsf.org). Since 2013, RSF has scored countries on their freedom of expression, based on responses to an online questionnaire that is distributed in 20 languages to a cross-section of media professionals, lawyers, and sociologists. The questionnaire includes 87 questions that address a variety of dimensions of free expression in a country, including pluralism, media independence, media environment, self-censorship, litigiousness, transparency, and the quality of the infrastructure that supports the production of news and information.



First time in Sea view

Review of Sea View Resort & Spa Koh Chang

●●●●○ Reviewed 6 April 2020

supawadeepho

1

It's good memory for us, First time for Sea view, it 's good view especially breakfast restuarant on top building. Good quality of food , quietly and good service mind But the price is expensive (high season).
More

Date of stay: February 2020

Figure 1. Screenshot of a Sample TripAdvisor Review

Accounting for Fake Reviews

Given that our primary analyses focus on review deletion, it is important to account for other systematic causes of review deletion. The most obvious alternative cause would be deletion by TripAdvisor. TripAdvisor actively seeks out and removes fake reviews; in fact, in 2020 alone, TripAdvisor reportedly removed 1 million fake reviews.[5] Accordingly, we also pre-processed our sample of reviews posted in 2021 to remove likely fake reviews, before undertaking our analysis of review deletion. To identify fake reviews on TripAdvisor, we leverage the approach introduced by Ott et al. (Ott et al. 2011; Ott et al. 2013). Ott and colleagues compiled a labeled sample of authentic and fake reviews for the TripAdvisor context and leveraged textual features to create a classifier. We re-use Ott et al.'s TripAdvisor sample for the same purpose. The data includes 1,600 reviews, amongst which 800 are truthful, and 800 are fake, i.e., the ground truth rate of deceptiveness is 50%. So, the data is balanced. We process the textual data, converting it to lowercase and removing stop words. Then, we vectorize the processed texts employing TF-IDF (term frequency-inverse document frequency). Finally, the resulting TF-IDF vector representations are fed to a Support Vector Machine (SVM) classifier. Our hyperparameters in the SVM are selected via 5-fold stratified cross-validation (CV) using the training sample. The accuracy of the resultant model is approximately 90% in the test sample (see Table 1 for the model performance).

Labels	Precision (%)	Recall (%)	F1-score (%)
Truthful	90	86	88
Deceptive	86	91	88

Table 1. SVM Model Performance

We also explored a deep learning-based method (BERT model). However, due to its poorer performance in precision/recall, we opted to retain the SVM. We applied our classifier to the entirety of our 2021 review sample (115,487), identifying 32,555 likely false reviews that had yet to be identified by TripAdvisor, at the time of our data collection. Fake reviews are deceptive reviews provided to mislead consumers in their purchase decision-making, often by reviewers with little or no experience with the products or services being reviewed. They could be also written by competitors to improve their reputation and drum up new business.

[5] https://www.forbes.com/sites/roberthart/2021/10/27/tripadvisor-took-down-nearly-1-million-fake-reviews-last-year/?sh=75e832682fa0

Excluding fake reviews, our censorship analyses proceeded with the remaining 82,932 reviews (Table 2). Among these 82,932 truthful reviews, approximately 1% were deleted by the end of 2021.

Review Censorship	Fake Review	Truthful Review	Total
No	32,240	82,098	114,338
Yes	315	834	1,149
Total	32,555	82,932	115,487

Table 2. Distribution of 2021 Sample of Reviews (Deceptive vs. Truthful)

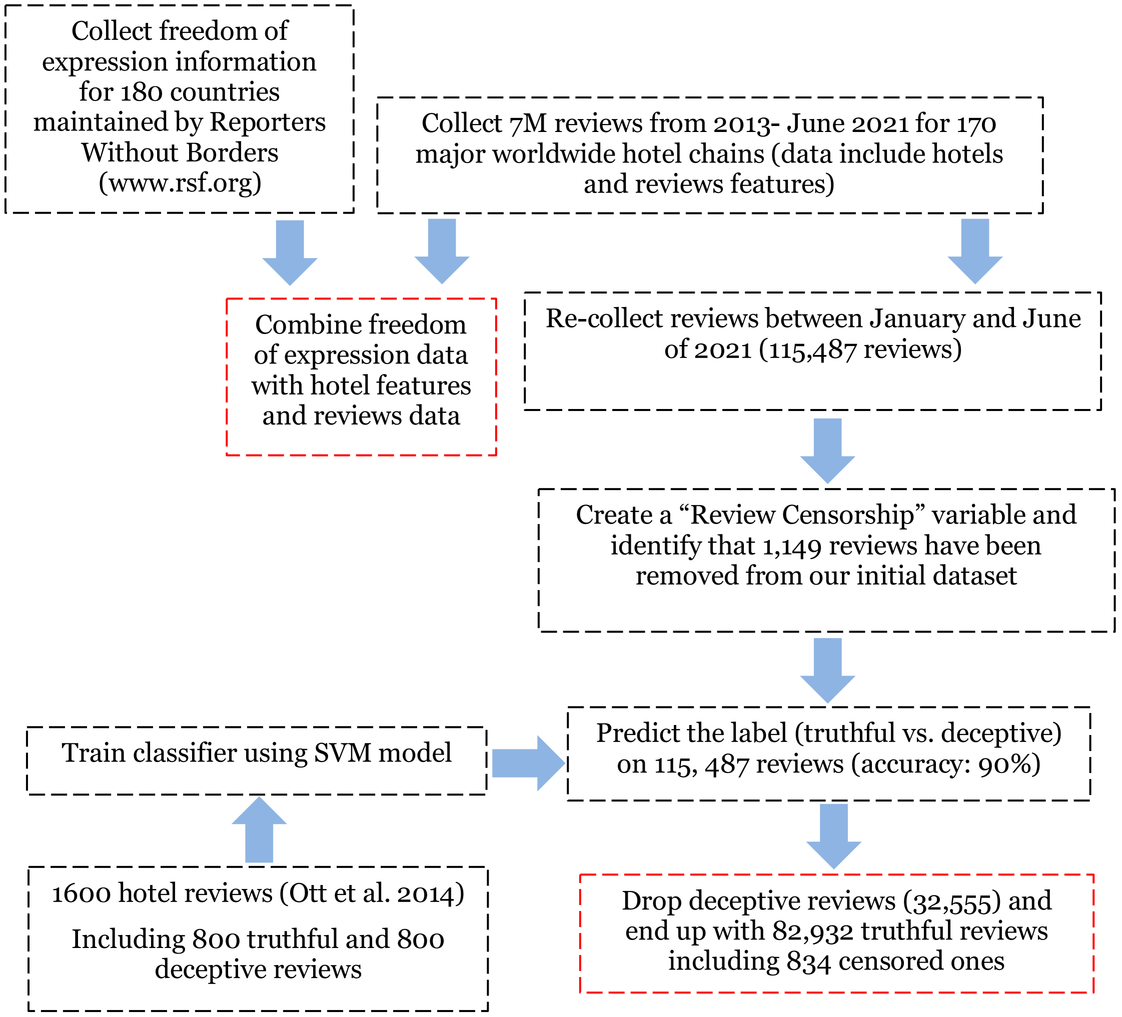

Collect freedom of expression information for 180 countries maintained by Reporters Without Borders (www.rsf.org)
Collect 7M reviews from 2013- June 2021 for 170 major worldwide hotel chains (data include hotels and reviews features)
Combine freedom of expression data with hotel features and reviews data
Re-collect reviews between January and June of 2021 (115,487 reviews)
Create a “Review Censorship” variable and identify that 1,149 reviews have been removed from our initial dataset
Train classifier using SVM model
Predict the label (truthful vs. deceptive) on 115, 487 reviews (accuracy: 90%)
1600 hotel reviews (Ott et al. 2014)
Including 800 truthful and 800 deceptive reviews
Drop deceptive reviews (32,555) and end up with 82,932 truthful reviews including 834 censored ones

Figure 2. Data Processing Framework

Lastly, Figure 2 shows all the steps of our data processing to construct the two datasets, outlined in red. A separate analysis will be examined on each of these datasets in the next sections.

Variables and Measures

Review Censorship: This measure was constructed for reviews in our sample that were originally posted in the first half of 2021. This is a binary variable, defined per review, which equals one if the original review was removed as of late 2021.

Rating Valence: Consumers rate hotels on TripAdvisor with a rating between 1 and 5 stars, with 5 being the most positive. This value is thus defined at the level of the review.

Review Sentiment: As noted above, we utilized VADER to operationalize textual sentiment based on the text of each review. VADER uses a combination of neutral or uncertain as well as how positive or negative the sentiment is. In addition, VADER calculates the compound, which is a continuous value between -1 (extremely negative) and +1 (extremely positive).

Review Delay: This measure reflects the number of months between the date a review was posted, and the recorded date of the hotel stay.

Review Volume: This variable is constructed by aggregating reviews up to the level of the hotel-year. This measure reflects the total number of reviews received by a given hotel, in a given year.

Restrictions of Free Expression (Score): As noted above, this country-level score is updated annually, per country, with possible values ranging from 0 to 100.

Restrictions on Freedom of Expression (Rank): In addition to the freedom of expression score, countries are also ranked between 1 and 180, with 1 being the best possible rank.

Table 3 presents the descriptive statistics for our key variables related to the review-level analysis of deletion (censorship), based on truthful reviews posted in the first half of 2021. Taking all reviews posted this 6 months and then removing all those labeled as likely to be fake, our final sample (labeled as truthful reviews) includes 82,932 observations (see Table 4 for the correlation matrix).

Variables	Obs.	Mean	Std. Dev.	Min	Max
Review Censorship	82,932	0.01	0.10	0.00	1.00
Rating Valence	82,932	3.94	1.49	1.00	5.00
Review Sentiment	82,932	0.62	0.60	-1.00	1.00

Table 3. Descriptive Statistics for 2021 Sample of Reviews (6 Months)

Variables	(1)	(2)	(3)
(1) Review Censorship	1.00		
(2) Rating Valence	0.01	1.00	
(3) Review Sentiment	0.01	0.77	1.00

Table 4. Correlation Matrix for 2021 Sample of Truthful Reviews (6 Months)

Similarly, Tables 5 and 6 present the descriptive statistics and a correlation matrix for the key variables associated with the sample of data used for our analysis of the relationship between review characteristics and local freedom of expression in a country. Note that the review volume measure in Table (5) is defined at the hotel-year level, while all the other variables are defined at the review level. For this reason, we do not include the review volume in the reported correlation matrix (Table 6).

Variables	Obs.	Mean	Std. Dev.	Min	Max
Rating Valence	6,278,248	3.91	1.22	1.00	5.00
Review Sentiment	6,278,248	.66	0.53	-1.00	1.00
Delay (months)	6,278,248	.43	1.37	0.00	83.00
Review Volume (hotel - year)	215,666	29.11	66.47	1.00	8,664.00
Restrictions on FoE (Score)	6,278,248	22.75	9.91	6.38	88.87
Restrictions on FoE (Rank)	6,278,248	44.99	32.87	1.00	180.00

Table 5. Descriptive Statistics (2013-2021) Reviews

Variables	(1)	(2)	(3)	(4)	(5)
(1) Rating Valence	1.00				
(2) Review Sentiment	0.66	1.00			
(3) Review Delay (month)	0.02	0.03	1.00		
(4) Restrictions on FoE (Score)	0.03	0.02	0.05	1.00	
(5) Restrictions on FoE (Rank)	0.03	0.02	0.06	0.95	1.00

Table 6. Matrix of Correlations (2013-2021) Reviews

Econometric Specification

In our analyses, we employ Ordinary Least Square (OLS) Regression. In Equation 1, the binary dependent variable is review censorship. Our regression thus reflects linear probability models (LPMs). We estimate the effects of rating valence and review sentiment on the probability of review deletion, while conditioning on hotel-chain, time (month), and reviewer fixed effects, i.e., μ_i, τ_t, and δ_j, respectively. In the equations, *i* indexes hotel chains, *j* indexes reviewers, and *t* indexes months.

$$[1]\ \text{Review Censorship}_{ijt} = \beta_0 + \beta_1 * \text{ rating valence}_{ijt}/\text{review sentiment}_{ijt} + \mu_i + \delta_j + \tau_t + \varepsilon_{ijt}$$

For the freedom of expression analyses, as noted earlier, the dependent variables include rating valence, sentiment, and delay, operationalized at the review level, and volume, operationalized at the hotel-year level. The key independent variable of interest is our country-level freedom of expression measure. Additionally, we again incorporate hotel-chain and reviewer fixed effects, as well as time (year) fixed effects, τ_t. Note that we rely on hotel-chain rather than hotel-fixed effects, as the freedom of expression measures exhibit relatively little variation year-over-year, making them highly collinear with the hotel-fixed effect. That said, hotel chains maintain a high degree of standardization in service delivery as part of their franchising agreements, despite being independently owned and operated. Accordingly, hotel-chain fixed effects can be expected to address much of the between-hotel variation in service quality and consumer experience. Please note that subscript j indexes reviewers. The freedom of speech score does not vary at the reviewer level; it varies only over time (t) and per hotel (i), depending on the hotel's location.

$$[2]\ \text{Rating Valence}_{ijt} = \beta_0 + \beta_1 * \text{ freedom of speech restriction score}_{it}\ /\text{rank}_{it} + \mu_i + \delta_j + \tau_t + \varepsilon_{ijt}$$

$$[3]\ \text{Sentiment}_{ijt} = \beta_0 + \beta_1 * \text{ freedom of speech restriction score}_{it}\ /\text{rank}_{it} + \mu_i + \delta_j + \tau_t + \varepsilon_{ijt}$$

$$[4]\ \text{Delay}_{ijt} = \beta_0 + \beta_1 * \text{ freedom of speech restriction score}_{it}\ /\text{rank}_{it} + \mu_i + \delta_j + \tau_t + \varepsilon_{ijt}$$

To construct review volume, we aggregated the data at the hotel-year level, indicating the number of reviews for each hotel and each year.

$$[5]\ \text{Volume}_{it} = \beta_0 + \beta_1 * \text{ freedom of speech restriction score}_{it} \,/\, \text{rank}_{it} + \mu_i + \tau_t + \varepsilon_{it}$$

Results

Review Censorship and Review Characteristics

Table 7 presents the results of our LPM regression, modeling the probability of review deletion as a function of review characteristics, hotel-chain fixed effects, month-fixed effects, and reviewer-fixed effects. Note that there are approximately 77,000 unique reviewers. As such, several thousand reviewers have posted repeatedly. As expressed in the table notes, the 5,598 reviews that contribute to our estimation are those that are retained after we account for reviewer, time, and hotel chain fixed effects. If a reviewer posts only once they will be dropped from the sample as repeated observations are required, per review, to implement the fixed effect regression. The same is true for time and hotel chain fixed effects.

From the analysis, we find that reviews are more likely to be deleted when they are negative in rating valence or sentiment, consistent with our expectations. These results collectively suggest that consumers are more likely to remove their reviews when they have expressed negative feedback, consistent with a mechanism of censorship. These results collectively support Hypotheses 1 and 2.

	(1)	(2)
VARIABLES	Review Censorship	Review Censorship
Rating Valence	-0.0025*** (0.001)	
Review Sentiment		-0.005*** (0.001)
Constant	0.0193*** (0.0024)	0.0126*** (0.001)
Hotel Chain FE	YES	YES
Reviewer FE	YES	YES
Month FE	YES	YES
Observations	5,598	5,598
R-squared	0.764	0.764
F stat (df)	16.21 (1, 71)	25.15 (1, 71)

*Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors in parentheses clustered by hotel chain, reviewer, and country; the 5,598 reviews that contribute to our estimation are those that are retained after we account for reviewer, time, and hotel chain fixed effects.*

Table 7. Effect of Rating Valence and Sentiment on Review Censorship

Freedom of Expression and Review Characteristics

We next evaluate our hypotheses about the relationships between local freedom of expression, and the average features of available reviews (over 8 years of data). We begin by considering the relationship between restrictions on freedom of expression and the valence of posted reviews. The results of these analyses are presented in Table 8. We find that when restrictions on freedom of expression are higher, the valence of reviews tends to be more positive. This is again consistent with an explanation of censorship, that negative reviews are diminished in those locations. We observe a similar result employing the rank measure of freedom of expression.

VARIABLES	(1) Rating Valence	(2) Rating Valence
Restrictions on FoE (Score)	0.00132** (0.0006)	
Restrictions on FoE (Rank)		0.000429** (0.0002)
Constant	3.935*** (0.0157)	3.946*** (0.0105)
Hotel Chain FE	YES	YES
Reviewer FE	YES	YES
Year FE	YES	YES
Observations	3,139,761	3,139,761
R-squared	0.485	0.485
F stat	5.306	5.913

*Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors in parentheses clustered by hotel chain, reviewer, and country; effective sample size is again reduced due to the inclusion of reviewer fixed effects, as a large fraction of reviewers contribute just one review within our sample.*

Table 8. Effect of FoE Restrictions (Score/ Rank) on Rating Valence

Next, we consider how freedom of expression impacts review sentiment. The results from Table 9 indicate restriction in free speech leads to more positive reviews and therefore higher textual sentiment which is also consistent with our hypothesis.

VARIABLES	(1) Review Sentiment	(2) Review Sentiment
Restrictions on FoE (Score)	0.000289* (0.000170)	
Restrictions on FoE (Rank)		0.000111** (0.00005)
Constant	0.704*** (0.005)	0.706*** (0.003)
Hotel Chain FE	YES	YES
Reviewer FE	YES	YES
Year FE	YES	YES
Observations	3,139,761	3,139,761
R-squared	0.406	0.406
F stat	2.879	4.501

*Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors in parentheses clustered by hotel chain, reviewer, and country.*

Table 9. Effect of Restrictions on FoE (Score/ Rank) on Review Sentiment

We then examine the speed with which consumers post their feedback. Table 10 indicates that higher restrictions on freedom of expression in the country where a hotel is located are positively associated with delays in review posting. This result provides support for our fifth hypothesis.

VARIABLES	(1) Review Delay	(2) Review Delay
Restrictions on FoE (Score)	0.00146*** (0.0005)	
Restrictions on FoE (Rank)		0.000328** (0.00016)
Constant	0.435*** (0.0111)	0.454*** (0.00681)
Hotel Chain FE	YES	YES
Reviewer FE	YES	YES
Year FE	YES	YES
Observations	3,139,761	3,139,761
R-squared	0.450	0.450
F stat	8.471	4.403

*Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors in parentheses clustered by hotel chain, reviewer, and country.*

Table 10. Effect of FoE Restrictions (Score/ Rank) on Review Delay

Finally, we look at the relationship between freedom of speech restriction and review volume. Since this analysis is performed at the hotel-year level, we aggregate over reviews, reducing the number of observations. According to the results in Table 11, greater restrictions on freedom of speech in a country are associated with fewer reviews accruing to local hotels. These findings are significant and are also consistent with Hypothesis 6.

VARIABLES	(1) Review Volume	(2) Review Volume
Restrictions on FoE (Score)	-0.357*** (0.0925)	
Restrictions on FoE (Rank)		-0.106*** (0.0351)
Constant	38.47*** (2.607)	35.20*** (2.068)
Hotel Chain FE	YES	YES
Year FE	YES	YES
Observations	215,664	215,664
R-squared	0.093	0.092
F stat	14.88	9.095

*Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors in parentheses clustered by hotel chain and country.*

Table 11. Effect of FoE Restrictions (Score/ Rank) on Review Volume

Discussion

The results of our study provide insights into a heretofore understudied issue in online review generation: review censorship. Online reviews play a crucial role in informing consumer purchase decisions. Review censorship can have very undesirable consequences for the market and consumer informedness, in a manner similar to fake reviews. Our estimates suggest that review censorship is prevalent; among all the truthful reviews posted in the first half of 2021, approximately 1% were removed within 6 months.

We have observed that review censorship is more likely to occur for low rating valence and negative sentiment reviews. Further, we find that censorship is more prevalent in locations where freedom of expression is reduced. In particular, we found that rating valence, sentiment, and time to posting are all positively associated with weaker freedom of expression in the location where a hotel is located. Further, we found that higher weaker freedom of expression in a location is associated with reduced reviewing volumes.

Our findings have important implications. First, we establish an upper bound on the prevalence of review censorship in the travel industry. Future work can expand on this analysis, to understand whether censorship varies substantially across other industries where ratings play a major role, e.g., restaurants, and physicians. Second, we demonstrate that censorship is associated with basic review characteristics. Future work can explore whether censorship also varies based on other factors, such as author characteristics. One major concern might be that certain segments of the consumer population are censored to a systematically greater degree, perhaps because they are more sensitive or wary of threats of legal action (e.g., immigrant populations or less wealthy individuals). Our work offers a framework for pursuing these questions.

Of course, this study is subject to several limitations. First, our sample is limited to TripAdvisor.com, which is unique in several respects that may limit generalizability. For example, TripAdvisor allows any individual to post reviews, without verification of consumption, a policy that differs from other platforms that host hotel ratings (e.g., Expedia.com). Future work might thus benefit from examining patterns of censorship on other review platforms, such as Yelp and Google Places.

Second, our study also relies exclusively on archival, observational data, which presents difficulties for exploring the nuances of censorship as an underlying mechanism, and its specific mode of occurrence. For example, it is difficult for us to distinguish between censorship arising from a retailer-provided incentive, like a discount (i.e., a bribe), and censorship deriving from a legal threat. Our initial exploration of user forum posts on TripAdvisor, Yelp, and Google Places suggests that both mechanisms are likely to be a play (consumers actively post questions indicating they are experiencing both types of activities). Accordingly, more work is needed to understand the prevalence of each phenomenon, as the proper policy responses to addressing each will differ. Future work might thus draw on alternative methodologies and data, e.g., surveying consumers to ask more detailed questions about experiences with censorship.

Third, our study is presently limited to the immediate act of censorship and deletion, without consideration of potential spillover effects on censored consumers later reviewing activity. Going forward, it would be interesting to examine individual-level reviewing behavior over time, to evaluate whether receiving threatening emails or phone calls from a business owner may depress negative feedback that consumers provide elsewhere.

Fourth, many hotels' bookings and flights got canceled due to COVID-19 and thus we may have fewer reviews. The pandemic started in March 2020 and continued for two years, and it appears to be endemic at this point. Importantly, the travel industry began to rebound in 2021. That said, the dynamics of reviewing may have differed during that period. Additionally, the data used to train the fake review detector is based on reviews from several reviewing platforms. As such, our results assume that reviews are homogenous across various platforms that cater to hotel travelers.

Last but not least, we also acknowledge that the result may be driven by our focus on English-language reviews, and the fact that FoE scores are systematically higher in English-speaking countries, thus the relationship between review volumes and FoE scores may be confounded by the prevalence of English speakers in a country.

References

Allard, T., Dunn, L. H., and White, K. 2020. "Negative Reviews, Positive Impact: Consumer Empathetic Responding to Unfair Word of Mouth," Journal of Marketing (84:4), pp. 86–108. (https://doi.org/10.1177/0022242920924389).

Ananthakrishnan, U. M., Li, B., and Smith, M. D. 2020. "A Tangled Web: Should Online Review Portals Display Fraudulent Reviews?," Information Systems Research (31:3), pp. 950–971. (https://doi.org/10.1287/ISRE.2020.0925).

Aral, S. 2014. "The Problem with Online Ratings," MIT Sloan Management Review (55:2), pp. 47–52. (https://sloanreview.mit.edu/media-download/29648/the-problem-with-online-ratings-2).

Calvert, C. 2018. "Gag Clauses and the Right to Gripe: The Consumer Fairness Act of 2016 & State Efforts to Protect Online Reviews from Contractual Censorship," Widener Law Review (24), p. 203.

Chevalier, J. A., and Mayzlin, D. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews," National Bureau of Economic Research (43:3), pp. 345–354. (https://doi.org/10.1016/j.thromres.2011.09.001).

Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., and Lee, L. 2009. "How Opinions Are Received by Online Communities: A Case Study on Amazon.Com Helpfulness Votes," WWW - Proceedings of the 18th International World Wide Web Conference, pp. 141–150. (https://doi.org/10.1145/1526709.1526729).

Dellarocas, C. 2003. "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," Management Science (49:10), pp. 1407–1424. (https://doi.org/10.1287/mnsc.49.10.1407.17308).

Fang, H., Zhang, J., Bao, Y., and Zhu, Q. 2013. "Towards Effective Online Review Systems in the Chinese Context: A Cross-Cultural Empirical Study," Electronic Commerce Research and Applications (12:3), pp. 208–220. (https://doi.org/10.1016/j.elerap.2013.03.001).

He, S., Hollenbeck, B., and Proserpio, D. 2022. "The Market for Fake Reviews," Marketing Science, pp. 1–26. (https://doi.org/10.2139/ssrn.3664992).

Hong, Y., Huang, N., Burtch, G., and Li, C. 2016. "Culture, Conformity, and Emotional Suppression in Online Reviews," Journal of the Association for Information Systems (17:11), pp. 737–758. (https://doi.org/10.17705/1jais.00443).

Lappas, T., Sabnis, G., and Valkanas, G. 2016. "The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry," Information Systems Research (27:4), pp. 940–961. (https://doi.org/10.1287/isre.2016.0674).

Lee, T. Y., and Bradlow, E. T. 2011. "Automated Marketing Research Using Online Customer Reviews," Journal of Marketing Research (48:5), pp. 881–894. (https://doi.org/10.1509/jmkr.48.5.881).

Lu, X., Ba, S., Huang, L., and Feng, Y. 2013. "Promotional Marketing or Word-of-Mouth? Evidence from Online Restaurant Reviews," Information Systems Research (24:3), pp. 596–612. (https://doi.org/10.1287/isre.1120.0454).

Luca, M. 2016. "Reviews, Reputation, and Revenue: The Case of Yelp.Com," Harvard Business School NOM Unit Working Paper No.12-016. (https://doi.org/10.2139/ssrn.1928601).

Luca, M., and Zervas, G. 2016. "Fake It till You Make It: Reputation, Competition, and Yelp Review Fraud," Management Science (62:12), pp. 3412–3427. (https://doi.org/10.1287/mnsc.2015.2304).

Mayzlin, D., Dover, Y., and Chevalier, J. 2014. "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," American Economic Review (Vol. 104). (https://doi.org/10.1257/aer.104.8.2421).

Min, H., Lim, Y., and Magnini, V. P. 2015. "Factors Affecting Customer Satisfaction in Responses to Negative Online Hotel Reviews: The Impact of Empathy, Paraphrasing, and Speed," Cornell Hospitality Quarterly (56:2), pp. 223–231. (https://doi.org/10.1177/1938965514560014).

Moutos, C. P., Verma, K., and Phelps, J. Y. 2020. "Negative Patient Reviews and Online Defamation: A Guide for the Obstetrician-Gynecologist," Obstetrics and Gynecology (136:6), pp. 1221–1226. (https://doi.org/10.1097/AOG.0000000000004150).

Mudambi, S. M., and Schuff, D. 2010. "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.Com," MIS Quarterly (34:1), pp. 185–200. (https://doi.org/10.2307/20721420).

Nakayama, M., and Wan, Y. 2019. "The Cultural Impact on Social Commerce: A Sentiment Analysis on Yelp Ethnic Restaurant Reviews," Information and Management (56:2), pp. 271–279. (https://doi.org/10.1016/j.im.2018.09.004).

Ott, M., Cardie, C., and Hancock, J. T. 2013. "Negative Deceptive Opinion Spam," North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference, pp. 497–501.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. 2011. "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (1), pp. 309–319.

Park, C., and Lee, T. M. 2009. "Information Direction, Website Reputation and EWOM Effect: A Moderating Role of Product Type," Journal of Business Research (62:1), pp. 61–67. (https://doi.org/10.1016/j.jbusres.2007.11.017).

Ponte, L. M. 2016. “Protecting Brand Image or Gaming the System? Consumer ‘Gag’ Contracts in an Age of Crowdsourced Ratings and Reviews.,” Willlim & Mary Business Law Reviewm & Mary Business Law Review (7:1). (https://doi.org/10.2139/ssrn.2579172).

Reimers, I., and Waldfogel, J. 2021. “Digitization and Pre-Purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings,” American Economic Review (111:6), pp. 944–1971. (https://doi.org/10.1257/AER.20200153).

Rose, M., and Blodgett, J. G. 2016. “Should Hotels Respond to Negative Online Reviews?,” Cornell Hospitality Quarterly (57:4), pp. 396–410. (https://doi.org/10.1177/1938965516632610).

Sen, S., and Lerman, D. 2007. “Why Are You Telling Me This? An Examination into Negative Consumer Reviews on the Web,” Journal of Interactive Marketing (21:4), pp. 76–94. (https://doi.org/10.1002/dir.20090).

Sleeper, M., Balebako, R., Das, S., McConahy, A. L., Wiese, J., and Cranor, L. F. 2013. “The Post That Wasn’t: Exploring Self-Censorship on Facebook,” Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, pp. 793–802. (https://doi.org/10.1145/2441776.2441865).

Stevens, J. L., Esmark Jones, C. L., and Breazeale, M. 2021. “Title Redacted: The Impact of Negative Online Review Censorship,” Journal of Product and Brand Management. (https://doi.org/10.1108/JPBM-04-2020-2877).

Stevens, J. L., Spaid, B. I., Breazeale, M., and Esmark Jones, C. L. 2018. “Timeliness, Transparency, and Trust: A Framework for Managing Online Customer Complaints,” Business Horizons (61:3), “Kelley School of Business, Indiana University,” pp. 375–384. (https://doi.org/10.1016/j.bushor.2018.01.007).

Unni, M. V. 2020. “Does Digital and Social Media Marketing Play a Major Role in Consumer Behaviour?,” International Journal of Research in Engineering, Science and Management (3:4), pp. 272–278.

Yin, D., Bond, S. D., and Zhang, H. 2014. “Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews,” MIS Quarterly (38:2), pp. 539–560.