"Not Some Trumped Up Beef": Assessing Credibility of Online Restaurant Reviews

Marina Kobayashi, Victoria Schwanda Sosik, David Huffaker

HAL Id: hal-01609412

https://inria.hal.science/hal-01609412

Submitted on 3 Oct 2017

"Not some trumped up beef": Assessing Credibility of Online Restaurant Reviews

Marina Kobayashi, Victoria Schwanda Sosik, and David Huffaker

Google, Inc., Mountain View, California
{kobayashim, victoriasosik, huffaker}@google.com

Abstract. Online reviews, or electronic word of mouth (eWOM), are an essential source of information for people making decisions about products and services, however they are also susceptible to abuses such as spamming and defamation. Therefore when making decisions, readers must determine if reviews are credible. Yet relatively little research has investigated how people make credibility judgments of online reviews. This paper presents quantitative and qualitative results from a survey of 1,979 respondents, showing that attributes of the reviewer and review content influence credibility ratings. Especially important for judging credibility is the level of detail in the review, whether or not it is balanced in sentiment, and whether the reviewer demonstrates expertise. Our findings contribute to the understanding of how people judge eWOM credibility, and we suggest how eWOM platforms can be designed to coach reviewers to write better reviews and present reviews in a manner that facilitates credibility judgments.

Keywords: eWOM · online review credibility · online review platforms

1   Introduction

Imagine you're in the market for a new computer or that you want to choose the perfect vacation destination. How do you decide the model that is best for your needs or pick from all the Caribbean islands? In these situations we have traditionally relied on word of mouth (WOM), or oral, person-to-person, non-commercial communication regarding a brand, product, or service [1]. Word of mouth has a huge impact on consumer behavior, holding more influence over people's choices, expectations, and attitudes than other types of information such as advertisements and neutral print sources [18], [35].

Much of this power comes from the perception that WOM is from other like-minded consumers who are not motivated to sell a product or service. However while it sometimes comes from this altruistic place, people also express opinions for reasons of self-enhancement, vengeance [12], or even satire (see Figure 1). Given these different motives, we must assess the credibility of others' opinions to decide how they will impact our own consumer decisions.
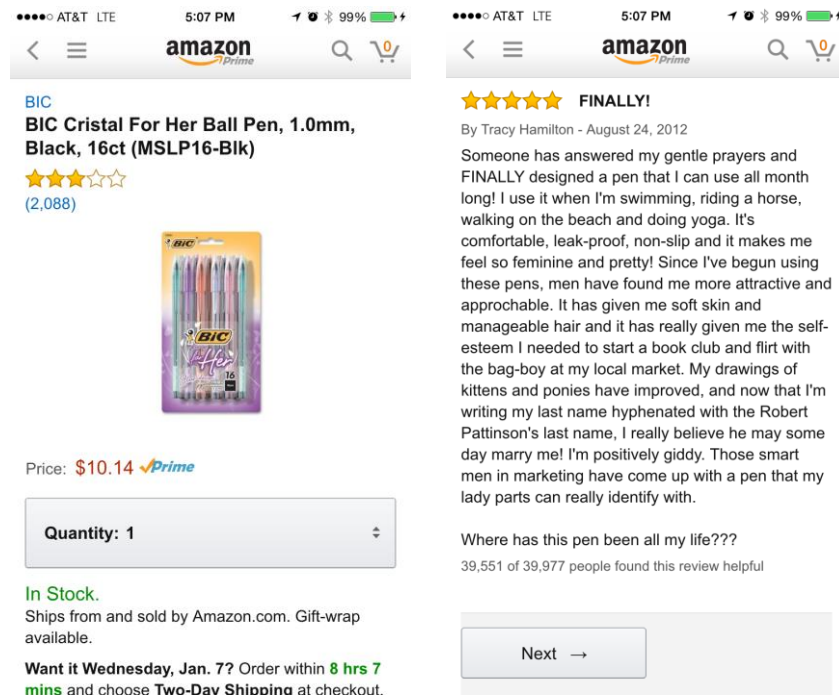
Fig. 1. Example of satirical review on Amazon.com

Beyond traditional ways to share opinions, use of the internet for WOM (eWOM) has become increasingly popular via online discussion forums, eCommerce sites (e.g., Amazon.com), and targeted opinion platforms (e.g., tripadvisor.com) [19]. These eWOM sources have extended the reach and altered the nature of WOM, allowing people to share opinions with a much larger audience of people whom they do not know personally. While such sites are great consumer resources, they've also been criticized as less credible than other information sources [16], [26]. Furthermore, they tend to lack standard means for assessing credibility, such as the identity of the information source [7].

In order to help people sort through the large volume of conflicting reviews that is common on eWOM platforms, and to do so in the absence of standard credibility signals, it is important to determine what other types of information may help people judge eWOM credibility. Yet there is very little work in this area of how people develop perceptions of credibility, especially in online environments [25]. Much of the prior work on credibility of online information has focused on traditional, institutionally created news media sites instead of user-generated content (e.g., [14]), which is often created by people lacking institutional authority. Some more recent work has investigated the credibility of user-generated content like blogs (e.g., [2]), wikis (e.g., [23]), and tweets (e.g., [8]), however with respect to verifiable, factual news events, not subjective opinions about products and services. Another line of work has developed algorithms to detect fake online reviews (e.g. [31]), however this work does

little to explain the process that people go through when making similar credibility judgments.

With the current research we strive to fill this gap and contribute an understanding of what types of information can help people judge the credibility of subjective eWOM reviews. These findings have implications for the design of eWOM platforms, suggesting ways such platforms can help reviewers write more helpful reviews and suggesting types of information that can be displayed alongside reviews to help readers make credibility judgments.

2   Related Work

2.1   Characteristics of Effective WOM

While the potential power of WOM is largely accepted (e.g., [1], [10]), it is not enough for WOM to simply be generated, it also needs to be judged as credible and recipients must be influenced by its content [39]. Despite the importance of understanding how WOM is received and the processes leading to its outcomes, the majority of research around WOM has focused on its generation [39]. However a smaller body of work has investigated factors impacting outcomes, finding that characteristics of the WOM's source are among the most important in explaining its influence [9]. For example the source's expertise and whether or not they are considered an opinion leader can impact a recipient's assessment [3], [15]. The tie strength between the source and recipient is also one of the strongest predictors of WOM influence, as closer friends have more personal knowledge of and a greater interest in the recipient [3], [5].

In addition to the source's characteristics, the content of the WOM message can also impact its influence. More vivid and strongly delivered messages--which can be conveyed by both message wording and body language—are more influential [39]. The valence of a message--whether it is positive or negative—also affects individuals' responses to WOM [21]. However findings around valence and influence have been inconsistent, with some work showing that positive messages have stronger impact [22] and other work showing the opposite [30].

2.2   Credibility Signals in eWOM

However as noted above, many of these factors impacting the assessment of WOM are absent in the case of eWOM [7]. Sources may be unknown to the recipient, may hide their identity behind pseudonyms, or may even post completely anonymously. eWOM is also written instead of oral, which may impact the interpretation of message content while removing accompanying body language and introducing the possibility for grammatical and spelling errors . As such, credibility judgments must be based on different factors in online eWOM environments than in traditional WOM.

Among the few studies that have begun to investigate this topic, source identity has been a focus. Studies have found that reviewers who disclose their identity information are viewed as more credible than those who don't [25], [40]. However the

identity manipulation in these studies provided extensive profile information such as name, location, age group, length of membership, picture, and a brief bio. While this type of information can help people judge similarity between themselves and the eWOM source, which can impact influence [4], people are not likely to include such detailed information in their profiles due to privacy concerns [29]. Prior work also shows that identity disclosures as simple as whether or not people use their real names can impact others' evaluation of online information such as news articles [13]. We therefore consider a simpler and more practical identity manipulation, focusing just on username:

> *RQ1: Is eWOM by reviewers who use their real names judged more credible than by those who use pseudonyms or post anonymously?*

In the absence of detailed source information, people often rely on other cognitive heuristics to assess the credibility of online information, such as reputation and endorsement from others [28], [37]. Online eWOM platforms have the opportunity to provide their audiences with cues to facilitate the use of these heuristics, yet what exactly these cues look like and whether they are effective has not be explored in prior work. Studies modeling the credibility of factual information in tweets show that information like the number of tweets people have made or the number of followers they have predicts credibility [8], [17], and these cues may likewise be useful for eWOM. In order to evaluate the efficacy of such signals, our second research question asks:

> *RQ2: How do reviewer status signals such as review or follower counts impact credibility judgments in eWOM?*

In addition to attributes related to source identity and reputation, valence has also been shown to impact assessments of credibility in eWOM. In one study, negative reviews were rated as more credible than positive ones, an effect that was strengthened when the reviewer disclosed their identity [25]. However given the mixed results of valence in traditional WOM (e.g., [22], [30]), we also investigate the influence of valence on eWOM credibility judgments in the current study. Furthermore, we extend prior work by considering more than just the extremes of positively and negatively valenced reviews. It is often the case that people express both positive and negative aspects of a product or service in a single, more balanced review and it is well established that balanced sentiment increases credibility in advertising [20]. Therefore we ask:

> *RQ3: How does eWOM valence (positive, negative, or balanced) influence credibility judgments?*
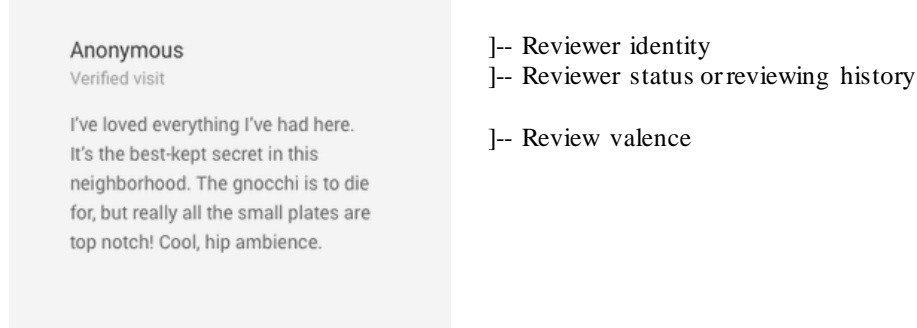
Anonymous
Verified visit

I've loved everything I've had here.
It's the best-kept secret in this
neighborhood. The gnocchi is to die
for, but really all the small plates are
top notch! Cool, hip ambience.

]-- Reviewer identity
]-- Reviewer status or reviewing history

]-- Review valence

Fig. 2. Example of stimuli that respondents were asked to rate

3 Methods

To answer our research questions, we recruited 1,979 U.S. respondents via Amazon Mechanical Turk (MTurk) to complete a survey, compensating them each with $1. MTurk has been used as a recruiting platform in several studies (e.g., [8], [22], [34]) and allows researchers to collect high-quality data from a more diverse population than the typical university student sample [6].

Each respondent was shown three restaurant review stimuli (see Figure 2) and asked to rate how credible they found each reviewer on a 7-point Likert scale. In order to understand how different signal variations impacted perceived credibility, we varied three elements of the review stimuli: reviewer identity, review valence and a UI signal related to the reviewer's status or reviewing history (see Table 1 for stimuli variations). Each respondent was presented with each of the three review sentiments in random order and reviewer identity and status variations were randomly combined with the review texts. When selecting real names to use in review stimuli, we generated a total of 118 names representing people from various nationalities and both genders (e.g., "Ellen Romano", "Hoang Kim"). We similarly included 6 different pseudonyms (e.g., "Natalie247", "DreamTeam4ever") in order to avoid effects due to a specific type of reviewer identity (e.g. American female who likes to cook).

We used linear mixed models to analyze how different review attributes affect credibility ratings, nesting credibility ratings within respondents. This method accounts for potential non-independence of observations since each respondent rated three different review stimuli. Note that the denominator degrees of freedom in linear mixed models are estimated using a Satterthwaite's approximation, which can yield non-integer degrees of freedom [38].

For the first review, we also asked respondents to describe in text why they rated the reviewer as they did. This allowed us both to verify that respondents were basing their rating on the stimuli presented and to understand what attributes of the review led to their credibility assessment. To understand factors that contributed to credibility judgments, we analyzed open-ended responses using open coding procedures [36].
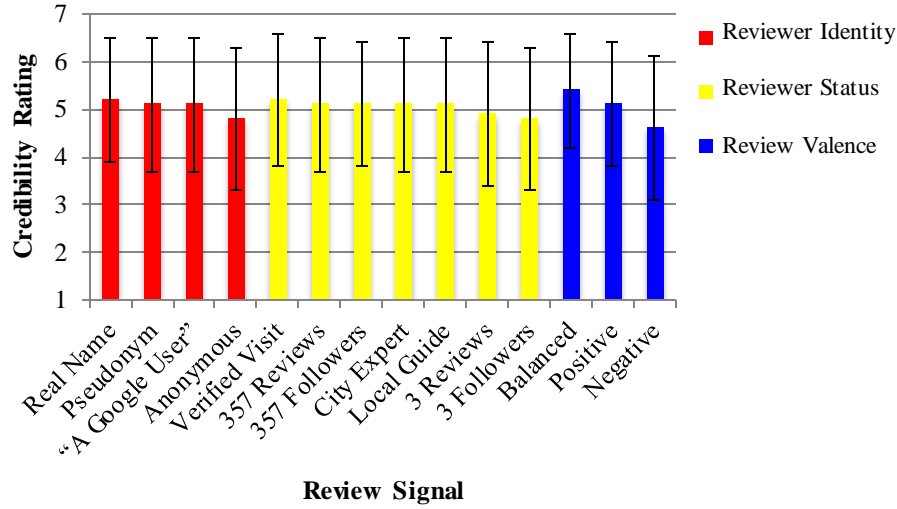
Table 1. Review stimuli variations.

Reviewer Identity	Reviewer Status	Review Valence
Real Name ex: *Ellen Romano, Hoang Kim*	**Stats: social, activity** *3 Followers* *357 Followers* *3 Reviews* *357 Reviews*	**Positive** *I've loved everything I've had here. It's the best-kept secret in this neighborhood. The gnocchi is to die for, but really all the small plates are top notch! Cool, hip ambience.*
In Lieu of Real Name *Anonymous* *A Google User*	**Labels: status, authentication** *City Expert* *Local Guide* *Verified Visit*	**Negative** *Overrated. We waited in line for what seemed like an eternity, only to get slow, impossible service and greasy, over-salted food. Will not be going back. Not worth the hype.*
Pseudonym ex: *Natalie247, DreamTeam4ever*		**Balanced** *Easy place to go for a quick bite. It gets pretty busy around dinner time, so you might want to consider getting takeout. All the dishes are reasonably priced and the portions are decent.*

Two researchers started by independently reviewing a random sample of 100 responses and iteratively generating a coding scheme. Note that in this scheme multiple codes can be assigned to a single response, if applicable. After arriving on a final scheme, we then individually reapplied it to the original set of 100 responses. We then resolved disagreements and calculated interrater reliability metrics, determining that another round of coding was necessary to reach acceptable reliability. We repeated this process with another set of 100 responses, this time reaching an average percent agreement of 97% (Krippendorff's alpha = 0.68). After resolving disagreements, we each coded an additional 150 responses, leading to a total of 500 coded responses (just over 25% of our data set).

4   Results

Together, our 1,979 respondents rated the credibility of 5,937 review stimuli. The average credibility rating across all stimuli was 5.0 on a 7-point Likert scale (SD = 1.39). Figure 3 presents the average credibility rating for different types of review attributes. We will discuss these differences in the following sections. While we model our three different types of attributes individually for ease of interpretation in the sections below, when including all of the attributes together, we find that our model explains 25% of variance in credibility ratings, as calculated using Ω_0^2 statistic for measuring explained variance in linear mixed models [41].

Review Signal

4.1 Effects of Review Attributes on Credibility Judgments

Reviewer Identity.

To answer RQ1, we analyzed how credibility ratings differed based on reviewer identity. We found that perceived credibility was significantly different between identity variations, $F[3, 4710.5] = 16.64$, $p < .001$ (see Table 2). Reviews from those who disclose their real name were judged as the most credible while reviews from those who posted anonymously were the least credible. We did not find significant differences between more traditional pseudonyms (e.g. Natalie247) and the "A Google User" identity.

Reviewer Status.

Our second research question investigates the impact of different types of signals about reviewer status on the judged credibility of reviews. We evaluated the effects of three different types of status signals. One signal acts to show the recipient how many reviews the reviewer has posted in the past, which can demonstrate expertise as a reviewer. The second type communicates the reviewer's reputation in the community by showing how many followers a reviewer has. The final type uses status labels like "Local Guide" to convey topic expertise.

We found that different types of status signals significantly impacted perceived credibility, $F[6, 5434.4] = 9.25$, $p < .001$ (see Table 2). Reviewers that were labeled with "Verified Visit" were judged the most credible, followed by those who demonstrated review expertise and reputation through their reviewing activities. Other types of status labels like "City Expert" and "Local Guide" were the next most credible. The least credible of all reviewers were those who were labeled as having less expertise or a small number of followers.

Review Valence.

To answer our third research question, we analyzed how different review valences influence perceived credibility ratings. We found review valence did significantly impact credibility, $F[2, 3956] = 223.56$, $p < .001$ (see Table 2) and balanced reviews were the most credible, followed by positive reviews then negative reviews. Based on prior work showing interactions between identity and valence in credibility judgments [25], we also modeled interactions between valence and status and valence and identity, however neither of these interactions were significant.

Table 2. Coefficients from three separate mixed models predicting credibility ratings based on different reviewer identities, reviewer status signals, and review valence.

Reviewer Identity[a]	Estimate (Std. Error)
Real Name	0.10 (0.05)*
Pseudonym	0.03 (0.05)
"A Google User"	--
Anonymous	-0.23 (0.05)***
Reviewer Status[b]	**Estimate (Std. Error)**
Verified Visit	0.42 (0.07)***
357 Reviews	0.34 (0.07)***
357 Followers	0.32 (0.07)***
City Expert	0.28 (0.07)***
Local Guide	0.27 (0.07)***
3 Reviews	0.13 (0.07)*
3 Followers	--
Review Valence[c]	**Estimate (Std. Error)**
Balanced	0.86 (0.04)***
Positive	0.49 (0.04)***
Negative	--

*$p < .05$, **$p < .01$, ***$p < .001$; a. Identities were dummy coded with "A Google User" as the reference identity; b. Status levels were dummy coded with "3 Followers" as the reference status; c. Valences were dummy coded with Negative as the reference valence*

4.2 Qualitative Factors Influencing Credibility Judgments

To gain a deeper understanding of what review attributes impact people's perceptions of review credibility, we also analyzed respondents' descriptions of their credibility judgments. Table 3 presents the data-driven codes that we arrived at along with example responses and the observed frequencies of each code. In line with prior work

on assessing credibility of WOM, our codebook includes two different types of codes, those relating to the reviewer themselves and those relating to the content of the message.

We also find that people described attributes of reviews that would both increase and decrease their credibility ratings. Exactly half of our response codes led to reported increases in credibility ratings (e.g. Displays Expertise, Relatable Reviewer, Many Followers), while the other half lead to reported decreases (e.g. Anonymous, Few Followers, Not Detailed). This finding is supported by looking at the average credibility rating for each code, which shows that all codes leading to reported decreases in credibility ratings had average ratings below the mean of 5.0, while all codes that were reported to increase credibility had average ratings above the mean.

We also find that some reasons to judge a review more or less credible are more frequently cited than others. The top three factors influencing credibility (by frequency) are how detailed, not detailed, or reasonable a review is. All of these codes were related to the review content instead of the reviewer, suggesting that respondents were more often focusing on review content than reviewers when making credibility judgments.

Table 3 also presents regression coefficients for a linear regression model predicting credibility rating based on the presence or absence of coded response factors. Note that we use linear regression here instead of linear mixed models since we are only modeling one rating per respondent. We find that our model explains 42% of the variance in credibility ratings, and the factors that most positively impact credibility ratings are the reviewer having many reviews, writing a reasonable review, and including much detail with balanced sentiment, respectively. The factors that most negatively impact credibility ratings are the use of pseudonyms, followed by having few followers displayed, the review not being detailed, and exaggerated or biased experiences.

Table 3. Frequencies of different factors reported to influence credibility judgments with respective average credibility ratings and regression coefficients for a model predicting credibility ratings.

Review Content	Example	Freq. (%)	Mean (SD)	Estimate (Std. Error)
Displays Expertise	*It sounds like she has been to a few restaurants in the area so she has some experience in how food should be. Not only that... but who orders gnocchi.. someone that knows food that is who.*	63 (12.6%)	5.7 (1.0)	0.59*** (0.15)
Actually Visited	*He mentioned a specific dish which tells me he was actually at the restaurant.*	59 (11.8%)	5.5 (1.0)	0.38* (0.16)
Relatable Reviewer	*This person seems like a friend who I would ask about restaurants.*	50 (10.0%)	5.7 (0.8)	0.65*** (0.17)
Many Fol-	*The person had hundreds of followers and*	18	5.1	-0.06

		Freq. (%)	Mean (SD)	Estimate (Std. Error)
lowers	*gave a review that seemed very plausible.*	(3.6%)	(1.1)	(0.26)
Anonymous	*The anonymous listing makes them less credible...I have no idea who they are.*	13 (2.6%)	4.3 (1.1)	-0.70* (0.31)
Not Relatable Reviewer	*Writing is mismatched forced. Like it's fake, no one says "hip" and why would you rave about a restaurant having small portions?*	12 (2.4%)	4.5 (1.3)	0.24 (0.32)
Many Reviews	*I gave him higher than I initially thought because of the 357 reviews he has given.*	10 (2.0%)	5.7 (0.9)	0.83* (0.35)
Few Followers	*They're semi credible. They only have 3 followers. I'd prefer to see more.*	8 (1.6%)	3.4 (1.1)	-1.44*** (0.39)
Few Reviews	*The person only has 3 reviews, so you can't really gauge whether or not she is credible.*	6 (1.2%)	4.0 (1.5)	-0.53 (0.45)
Pseudonym	*Scrubs999 does not sound like a credible reviewer, so I would be unlikely to believe what they are telling me. If the signed their name or even first name and last initial I would think this would be more credible.*	4 (0.8%)	2.8 (1.2)	-1.81** (0.55)

Review Content	Example	Freq. (%)	Mean (SD)	Estimate (Std. Error)
Detailed	*It seems quite thorough. Succinctly stated his experience from beginning to end.*	115 (23.0%)	5.7 (1.0)	0.77*** (0.13)
Not Detailed	*They gave no specifics about the restaurant such as recommended dishes. It was too general.*	107 (21.4%)	3.9 (1.4)	-0.98*** (0.13)
Reasonable Review	*They listed several reasons for the rating they give. The reasons sounded realistic and not some trumped up beef*	78 (15.6%)	5.7 (1.0)	0.79*** (0.14)
Biased Experience	*It sounds like emotion got in the way of a fact-based or purely observational review.*	48 (9.6%)	3.7 (1.4)	-0.74*** (0.18)
Well written	*The reviewer wrote in full sentences using good language/grammar.*	44 (8.8%)	5.5 (1.2)	0.29 (0.18)
Not Exaggerated	*There's nothing dramatic about the person's statement. I don't see exclamation marks saying it's sooo fantastic!!!*	37 (7.4%)	5.5 (1.0)	0.36 (0.19)
Exaggeration	*Instead of giving a thorough review of the food, the reviewer attacked the establishment with over-exaggerations.*	21 (4.2%)	3.6 (1.3)	-0.80** (0.03)

Balanced Sentiment	*The review doesn't seem over the top, one way or the other. Just a balanced view of things.*	19 (3.8%)	5.8 (0.9)	0.60* (0.26)
Too Positive	*The review is overly positive, which can raise some red flags.*	12 (2.4%)	4.2 (1.3)	-0.04 (0.33)
Poorly Written	*It is written in very basic language, something I expect to see a teenager write or someone with no experience.*	6 (1.2%)	4.0 (2.4)	-0.46 (0.45)

5   Discussion

In summary, by showing people different types of review stimuli, we find that reviewer identity, reviewer status, and review valence all impact credibility judgments of online restaurant reviews. People find reviewers who use their real names more credible than those who post anonymously, which extends prior work by showing that differences in credibility can be seen even when just varying the reviewers' name as opposed to requiring a more drastic identity manipulation like hiding or revealing detailed profiles [25]. We also find that people judge reviewers with a labeled status signaling their expertise or reputation in a community as more credible than those lacking this signal. This finding provides empirical support for work suggesting different cues and heuristics that people use to judge the credibility of online information [28], [37]. We find support for these heuristics in the case of eWOM specifically, and inspired by work on judging factual information in tweets [8] we also suggest concrete signals to facilitate heuristic credibility judgment in eWOM. Signaling domain expertise by showing that a reviewer had actually visited a restaurant led to the highest credibility ratings, followed by signals that show a reviewer has posted many reviews and has many followers. Finally, we find that balanced reviews were judged the most credible, followed by positive and then negative reviews. Our finding contributes to the conflicting body of work evaluating how review valence impacts credibility, because unlike prior work [25], we find that positive eWOM is judged more credible than negative eWOM. This suggests that in an online environment, the impact of review valence is just as unclear as in traditional WOM [22], [30] and further work is needed to better understand when and how review valence impacts credibility judgments. Although consistent with longstanding research in advertising [20], we find that a balanced review was judged to be even more credible than either a positive or negative review, and suggest that future work consider this case in addition to the typically studied positive/negative valence dichotomy.

Our qualitative analysis of people's descriptions detailing why they judged credibility in the way that they did also provides a deeper understanding of how review attributes impact eWOM credibility ratings. People reported attending to each of the three different types of signals that we manipulated: reviewer identity (codes: *Anonymous*, *Pseudonym*), reviewer status (codes: *Displays Expertise*, *Actually Visited*, *Many Followers*, *Many Reviews*, *Few Reviews*, *Few Followers*), and review valence

(codes: *Biased Experience, Too Positive, Balanced Sentiment*). They also used other criteria to judge credibility of reviews, considering how relatable (or not) a reviewer seemed, how detailed (or not) a review was, how reasonable the review content felt, whether or not it was exaggerated, and whether or not it was well written. We find that these codes are consistent with other data-driven heuristics that people base credibility judgments on, but are more detailed and apply specifically to the case of eWOM instead of online information more generally [28], [37]. For example, in eWOM *authority* and *reputation* can be gained by displaying expertise about a particular topic (e.g. demonstrating via status signal or detailed review content that one has actually visited a restaurant) and by showing that one has an extensive review history. Having a large number of followers also suggests that one has been *endorsed* by others, leading to increased credibility. People's consideration of whether the review seemed reasonable to them is also in line with the *expectancy violation* heuristic, stating that people find information more credible when it is consistent with their expectations [28]. Additionally, people talked about how relatable the reviewer was, which follows from [28]'s *persuasive intent* heuristic, or the fact that people find commercial information less credible than information from someone like themselves. Finally, we find further evidence that in the case of eWOM, presentation attributes like the quality of writing in a review can impact its credibility rating [27] just as presentation attributes like body language can impact judgments of traditional WOM [39].

5.1   Relative Importance of Reviewer and Review Attributes

Our work also contributes an understanding of the relative importance of different types of signals in eWOM. Studies of WOM have prioritized attributes of the reviewer in making credibility judgments [3], [5], [15], however we find that attributes of the review tended to be the most influential. Examining model coefficients shows that differences in review valence most strongly predicted credibility ratings, beyond reviewer identity or status. The most commonly cited explanations for credibility ratings in our open-ended responses were also attributes of the review, not the reviewer, suggesting that people were consciously paying more attention to the review content. Furthermore, five of the top eight explanations that actually impacted credibility ratings were related to the review content. It is therefore likely that while signals as to reviewer identity, expertise, and reputation are still useful, in the absence of first-hand knowledge of these attributes (such as is the case in eWOM), people focus more on the content itself when judging credibility.

5.2   Implications for Designing eWOM Platforms

Our work has implications for the design of platforms that solicit and present eWOM content, like Amazon.com and tripadvisor.com. Based on our findings showing how important the content of a review is in judging credibility, such sites should provide people with more guidance during the review creation process. For example, review valence can be automatically determined [33] as people write a review, and if a review appears overly positive or negative, the site could make suggestions that the

reviewer try to include a more balanced account of their experience with the product or service. Likewise they could be given tips on writing a more detailed review if the review appears to be too short or lacking in descriptive words. This type of reviewer coaching could facilitate the creation of more helpful and eventually more influential eWOM.

It also suggests types of signals that should be presented alongside reviews in order to support people in judging credibility. Selecting and verifying types of information that can signal a reviewers' domain expertise is important, as we saw that the "Verified Visit" signal most positively impacted credibility judgments of restaurant reviews. These platforms do not always have access to such information, but often they do by way of purchase records or records of online restaurant reservations. When this information is lacking, it can still be useful to show a record of how many reviews a reviewer has written or how respected they are in the community, via the number of people that have read and/or subscribe to their reviews and/or find them helpful.

5.3 Limitations

While we make several contributions to the understanding of how credibility is signaled and assessed in eWOM, our work is not without limitations. We used MTurk as a recruiting platform, and while this allowed us to quickly collect high-quality data, studies have found that those on MTurk are more likely female than male, averaging 36 years of age and more highly educated than the general US population [31]. Our methods also asked respondents to evaluate the credibility of individual reviews, while reviews more often appear amongst others in eWOM platforms. This allowed us to isolate specific variations of different types of review signals, but does not consider the effect that groups of reviews have on each other, as has been studied in other work [10]. Finally, some of our model coefficients were small, representing significant but not substantial changes in credibility ratings; this is particularly true for the model predicting reviewer identity. Therefore even though we find reviewers who use their real names to be most credible, we're hesitant to suggest that eWOM platforms require all users to use their real names due to potential privacy concerns [29] and related drops in contribution rates for not all that great of an increase in review credibility.

6 Conclusion

Our work contributes to the understudied topic of how review attributes impact judgments of eWOM credibility, focusing on attributes of the review and of the reviewer. We also present a deeper qualitative analysis of factors that respondents describe focusing on when making their credibility judgments. We find that many of the same factors important in traditional WOM assessments are also important in eWOM judgments, however they are signaled differently. We also see that people may place more emphasis on the review content itself as opposed to the reviewer when judging eWOM. This work has implications for designing eWOM platforms that both coach reviewers to write better reviews and present reviews to recipients in a manner that

facilitates credibility judgments. Future work should continue to investigate the impact of review valence on eWOM credibility as well as evaluate a more exhaustive list of concrete signals that can help people to make use of cognitive heuristics in assessing online review credibility.

Acknowledgements

References

1. Arndt, J. (1967) Role of product-related conversations in the diffusion of a new product. *Journal of Marketing Research 4*, 291−295.
2. Banning, S. A., & Sweetser, K. D. (2007). How much do they think it affects them and whom do they believe? Comparing the third-person effect and credibility of blogs and traditional media. *Communication Quarterly, 55*, 451–466.
3. Bansal, H.S. and Voyer, P.A. (2000), Word-of-mouth processes within a service purchase decision context. *Journal of Service Research, 3*, 2, 166-77.
4. Bart, Y., Shankar, V., Sultan, F., & Urban, G. L. (2005). Are the drivers and role of online trust the same for all web sites and consumers? A large-scale exploratory empirical study. *Journal of Marketing 69*, 133-152.
5. Brown, J. J., & Reingen, P. H. (1987). Social ties and word-of-mouth referral behavior. *Journal of Consumer Research, 14*, 3, 350−362.
6. Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data*?, Perspectives on Psychological Science, 6*, 1, 3–5.
7. Callister, T. A., Jr. (2000). Media literacy: On-ramp to the literacy of the 21st century or cul-de-sac on the information superhighway. *Advances in Reading/Language Research, 7*, 403−420.
8. Castillo, C., Mendoza, M., & Poblete, B. (2011). Information Credibility on Twitter. In *Proceedings of WWW 2011*, 675–684.
9. Dichter, E. (1966). How word of mouth advertising works. *Harvard Business Review, 44*, 6, 147-160.
10. Doh, S.-J., & Hwang, J.-S. (2009). How Consumers Evaluate eWOM (Electronic Word-of-Mouth) Messages. *CyberPsychology & Behavior, 12*, 2, 193–197.
11. Dye, R. (2000). The buzz on buzz. *Harvard Business Review, 78*, 139–146.
12. Engel, J. F., Blackwell, R. D., & Miniard, P. W. (1993). *Consumer Behavior* (8th ed.). Fort Worth: Dryden Press.
13. Fitch, J. W., & Cromwell, R. L. (2001). Evaluating Internet Resources: Identity, Affiliation, and Cognitive Authority in a Networked World. *Journal of the American Society for Information Science and Technology, 52*, 6, 499–507.
14. Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of Internet information credibility. *Journalism and Mass Communication Quarterly, 77*, 515–540.

15. Gilly, M. C., Graham, J. L., Wolfinbarger, M. F. & Yale, L. J. (1998). "A dyadic study of interpersonal information search", *Journal of the Academy of Marketing Science, 26,* 2, 83–100.

16. Gretzel U., Fesenmaier D. R., & O'Leary, J. T. (2006). The transformation of consumer behaviour. In: Buhalis D and Costa C (eds) *Tourism Business Frontiers*. Burlington, MA: Elsevier/Butterworth-Heinemann., pp. 9–18.

17. Gupta, A., & Kumaraguru, P. (2012). Credibility Ranking of Tweets during High Impact Events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media* (PSOSM '12).

18. Herr, P.M., Kardes, F.R. and Kim, J. (1991) Effects of word-of-mouth and product attribute information on persuasion: an accessibility-diagnosticity perspective. *Journal of Consumer Research, 17,* 454−62.

19. Horrigan, J. (2008, January 7, 2015). The Internet and consumer choice. Available: http://www.pewinternet.org/Reports/2008/The-Internet-and-Consumer-Choice.aspx

20. Hunt, J. M., & Smith, M. F. (1987). The Persuasive Impact of Two-Sided Selling Appeals for an Unknown Brand Name. *Journal of the Academy of Marketing Science, 15,* 1, 11−18.

21. Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology, 64,* 4, 349–371.

22. Kittur, A., Chi, E. H., & Suh, B.. (2008). Crowdsourcing User Studies With Mechanical Turk. In *Proceedings of the ACM Conference on Human-factors in Computing Systems* (CHI 2008), 453–456.

23. Kittur, A., Suh, B., & Chi, E. H. (2008). Can You Ever Trust a Wiki? Impacting Perceived Trustworthiness in Wikipedia. In *Proceedings of Computer-Supported Cooperative Work* (CSCW 2008), pp. 477–480.

24. Klayman, J., & Ha, Y.-W. (1987.) Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 2, 211–228.

25. Kusumasondjaja, S., Shanka, T., & Marchegiani, C. (2012). Credibility of online reviews and initial trust: The roles of reviewer's identity and review valence. *Journal of Vacation Marketing, 18,* 3, 185–195.

26. Litvin S. W., Goldsmith, R. E., & Pan, B. (2008) Electronic word-of-mouth in hospitality and tourism management. *Tourism Management, 29,* 3, 458–468.

27. Mackiewicz, J. (2008). Reviewer motivations, bias, and credibility in online reviews. In S. Kelsey & K. S. Amant, (Eds.), *Handbook of Research on Computer Mediated Communication,* p.p. 252–266. IGI Global.

28. Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication, 60,* 3, 413−439.

29. Milne, G. R., Rohm, A. J., & Bahl, S. (2004). Consumers' protection of online privacy and identity. *Journal of Consumer Affairs, 38,* 2, 217–232.

30. Mizerski, R. W. (1982). An attribution explanation of the disproportionate influence of unfavorable information. *Journal of Consumer Research, 9,* 301–310.

31. Ott, M., Cardie, C., & Hancock, J. (2012). Estimating the Prevalence of Deception in Online Review Communities. In *Proceedings of WWW 2012*.

32. Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment & Decision Making, 5,* 5, 411−9.

33. Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count (LIWC)* [Computer software]. Mahwah, NJ: Erlbaum.

34. Shay, R., Ion, I., Reeder, R. W., & Consolvo, S. (2011). My religious aunt asked why I was trying to sell her viagra": Experiences with account hijacking. In *Proceedings of CHI 2014,* 2657−2666.

35. Sheth, J.N. (1971) Word of mouth in low risk innovations. *Journal of Advertising Research, 11*, 15−18.

36. Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.

37. Sundar, S. S. (2008). The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. *Digital Media, Youth, and Credibility*, 73−100.

38. SPSS Technical Report-Linear Effects Mixed Modeling. http://www.spss.ch/upload/1107355943_LinearMixedEffectsModelling.pdf.

39. Sweeney, J. C., Soutar, G. N., & Mazzarol, T. (2007). Factors influencing word of mouth effectiveness: receiver perspectives. *European Journal of Marketing, 42*, 3/4, 344−364.

40. Xie, H., Miao, L., K, P.-J., & Lee. B.-Y. (2011). Consumers' responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition. *International Journal of Hospitality Management, 30*, 178−183.

41. Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine, 22*, 22, 3527−3541.