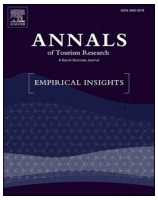
Annals of Tourism Research Empirical Insights 2 (2021) 100027

Contents lists available at ScienceDirect

Annals of Tourism Research Empirical Insights

journal homepage: www.sciencedirect.com/journal/annals-of-tourism-research-empirical-insights



Deepfake consumer reviews in tourism: Preliminary findings

Aarni Tuomi

Hospitality Business Competence Area, Haaga-Helia University of Applied Sciences, Pajuniityntie 11, 00320 Helsinki, Finland

ARTICLE INFO

Keywords:
Deepfake
Online review
Artificial intelligence
AI
Natural language processing
Gpt-2

1. Introduction

Electronic word-of-mouth has become a key part of decision-making in the digital age (Karakaya & Barnes, 2010), whereby particularly in tourism, user generated content (UGC) such as online consumer reviews have been found to play a significant role e.g. in travel companies' reputation management, the overall tourist customer journey, and in boosting hotel room sales (Baka, 2016; Yachin, 2018; Ye, Law, & Gu, 2009). Along other implications for tourism management, the increase in online review activity has given rise to the phenomenon of fake reviews (Yoo & Gretzel, 2009). Often written to promote (or demote) tourism businesses through "digital deception" (Choi et al., 2016), fake reviews mislead readers and in doing so may impact e.g. brand image. A hotel might for instance benefit from posting or soliciting fraudulent positive reviews about its own properties and negative reviews about its competitors' properties (Mayzlin, Dover, & Chevalier, 2014), causing issues for consumers, tourism businesses, as well as employees who have to read and reply to reviews. Hoping to understand the scale of the phenomenon better, Luca and Zervas (2016) estimated that out of all reviews on Yelp 10–20% are fake, while in a similar vein, in their longitudinal analysis of fraudulent activity on tourism review site TripAdvisor, Harris (2018) found strong evidence of fake reviews on the platform.

While fake reviews have traditionally been made-up and written by people, recent advances in artificial intelligence offer powerful new tools for online spin doctors. The study of "deepfakes", broadly understood to mean any type of content generated automatically by a machine learning system, is a booming area of research (Westerlund, 2019). Given the importance of consumer reviews in tourism, the rest of this paper discusses findings from two preliminary studies which test the feasibility of contemporary machine learning techniques to generate believable fake reviews in tourism contexts and explore the subsequent implications of doing so.

2. Study 1: Human- vs. computer-generated reviews

In Study 1, a total of 10 fake restaurant reviews were generated using OpenAI's natural language generator GPT-2. GPT-2 is an open-source natural language processing model that has been trained on eight million text documents scraped from the internet. Using a combination of four tokens: "this", "restaurant", "café", and "bar", GPT-2 was given the start of the sentence, e.g. "this restaurant", while the following 20–30 words were generated randomly. The script used was supervised, whereby every few words the researcher was prompted with a choice of possible follow-up words. In these instances the words that followed the desired narrative (i.e. restaurant review) were chosen. The resulting computer-generated review data were complemented by randomly scraping 10 human-authored restaurant reviews from TripAdvisor. Finally, a randomized between subject choice experiment was designed, whereby tourism employees ($n = 100$) who, as part of their job, read and reply to consumer reviews were asked to evaluate whether they thought the reviews they were presented with were written by a human or were generated by a computer. Descriptive statistics were calculated and the contents of the most human-like / not human-like computer-generated reviews were qualitatively analyzed to identify any recurring patterns.

Altogether 1000 evaluations were given; of these, 44% were found to be incorrect. Out of the 10 fake reviews generated, three were found to be particularly convincing, with an incorrect label being allocated in

E-mail address: aarni.tuomi@haaga-helia.fi.

https://doi.org/10.1016/j.annale.2021.100027
Received 17 May 2021; Received in revised form 29 June 2021; Accepted 29 July 2021
Available online 10 August 2021

85% of cases. Further analysis of the most human-like and the least human-like computer-generated reviews revealed that reviews which were not overly positive and included critical statements (e.g. "the food is not the best but it is still delicious"), as well reviews which included a call to action (e.g. "I'd definitely recommend stopping by") were perceived as particularly human-like. On the other hand, reviews that overused adjectives, as well as reviews which simply listed menu items or were very formal or to-the-point (e.g. "the eatery is located on the third floor of the building") were perceived as particularly machine-like.

3. Study 2: "Humanness" of computer-generated reviews

To explore the human- or machine-likeness of computer-generated reviews further, Study 2 sought to understand the degree to which the reviews' sentiment (positive, negative, or mixed) plays a role in how convincing (i.e. human-like) a review is perceived. Using the same process as in Study 1, a total of 15 restaurant reviews (of which 5 were positive, 5 negative, and 5 mixed) were generated with GPT-2. The order of the reviews was randomized, and following a purposive sampling approach, a different set of tourism employees ($n = 32$) were asked to evaluate the human- or machine-likeness of the review on a 5-point bipolar Likert scale (−2 Very Machine-Like, 2 Very Human-Like). Participants were also asked to highlight any words, phrases, or expressions they considered particularly human- or machine-like. Again, descriptive statistics were calculated and the highlighted content of the reviews was qualitatively analyzed.

The mean of evaluations was 0.48 (sd: 0.86), indicating a skew towards human-likeness. Overall, reviews with a negative or mixed sentiment were perceived as more human-like than reviews with a positive sentiment [means: 0.97 (neg.), 0.69 (mix.), −0.66 (pos.)]. Features highlighted as particularly machine-like fell into four categories: 1) repetition, i.e. using the same word or a limited number of words repeatedly, 2) using multiple adjectives to describe some element of the dining experience, 3) focusing on something that is not related to the core offering e.g. the location, appearance of staff or reputation of the venue, and 4) words and phrases that entailed an assumed hidden agenda, e.g. to convince the reader to visit the establishment (e.g. "if you're in the area"). Complementing these, features highlighted as particularly human-like also fell into four categories: 1) swearing, 2) the use of superlatives (e.g. best, worst, cheapest), 3) playing with words or using creative expressions, and 4) using personal pronouns and writing in first person.

4. Implications & future research

Recent tourism literature has highlighted the importance of conceptualizing the impacts of "fake news" on tourism (Fedeli, 2021). This research note extends these discussions by calling for more attention to the phenomenon of "deepfakes", particularly deepfake online consumer reviews and their impacts on tourism management theory and practice (Juuti, Sun, Mori, & Asokan, 2018). Previous studies on human-authored fake reviews in tourism have drawn on theories ranging from deception theory (Yoo & Gretzel, 2009) to source credibility theory (Ayeh, Au, & Law, 2013), among others. In their work, Ayeh et al. (2013) for example highlight the impact of homophily on credibility perceptions, whereby reviews written and read by like-minded people might be perceived as particularly credible. Further, seeking to mitigate the impacts of human-authored fake reviews, tourism scholars have suggested a myriad of strategies for identifying and dealing with fraudulent UGC. There is consensus that attention should be paid to both the profile of the review-giver as well as the actual contents of the review, including e.g. time of registration, number of reviews given, the frequency and extremity of reviewing activity, the lexicon used, as well as the comprehensiveness of the review (Liu & Hu, 2021; Luca & Zervas, 2016; O'Connor, 2008; Yoo & Gretzel, 2009).

Illustrated by the two preliminary studies presented herewith, strategies for identifying deepfake online consumer reviews seem to be well in line with previous strategies developed for dealing with human-authored fake reviews, perhaps with particular emphasis on the lexicon, sentiment, and the overall comprehensiveness of the review. As discussed by Westerlund (2019), deepfakes approximate content, whereby the output is a slightly altered version of the input. Because of this, computer-generated text tends to be less coherent and more along stream of consciousness writing, with incomplete ideas and the narrative taking illogical turns at times. Tourism managers should therefore pay particular attention to the comprehensiveness of the review as a whole. Further, in terms of broader impacts, what distinguishes deepfake reviews from human-authored fake reviews is the potential volume of fraudulent content (Diresta, 2020), whereby generating deepfake text seems much less resource-intensive than soliciting human-authored fake reviews (Mayzlin, 2006). At the extreme end, this may lead to situations where tourism review sites get flooded with convincing, low-cost computer-generated content which in turn influences decision-making e.g. through the so-called majority illusion (Lerman, Yan, & Yu, 2016). Given how COVID-19 has exacerbated the collective move to digital (Soto-Acosta, 2020), this research note seeks to demonstrate the implications of computer-generated fake reviews for tourism management. In doing so, the paper provides tourism scholars preliminary insight into how deepfake online reviews influence tourism management, including the kinds of features that make a given narrative particularly "human- or machine-like". Future research should continue this line of inquiry by exploring strategies for detecting, moderating, and replying to computer-generated reviews in tourism. In particular, attention should be paid to exploring impacts of computer-generated reviews across different review platforms (Xiang, Du, Ma, & Fan, 2017), use-contexts (e.g. accommodation; multinational corporation), user characteristics (e.g. age; experience on the job), as well as lexical differences (e.g. formal language; use of emoticons) (Huang, Chang, Bilgihan, & Okumus, 2020).

References

Ayeh, J. K., Au, N., & Law, R. (2013). "Do we believe TripAdvisor?" examining credibility perceptions and online travelers' attitude towards using user-generated content. *Journal of Travel Research, 52*(4), 437–452.

Baka, V. (2016). The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector. *Tourism Management, 53*, 148–162.

Choi, S., Mattila, A., Van Hoof, H., & Quadri-Felitti, D. (2016). The Role of Power and Incentives in Inducing Fake Reviews in the Tourism Industry. *Journal of Travel Research, 56*(8), 975–987. https://doi.org/10.1177/0047287516677168

Diresta, R. (2020). AI-generated text is the scariest deepfake of all. https://www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/. (Accessed 17 May 2021).

Fedeli, G. (2021). "Fake news" meets tourism: A proposed research agenda. In *Annals of tourism research* (p. 80). https://doi.org/10.1016/j.annals.2019.02.002

Harris, C. (2018). Decomposing TripAdvisor: Detecting potentially fraudulent hotel reviews in the era of big data. In *2018 IEEE international conference on big knowledge*. https://doi.org/10.1109/ICBK.2018.00040

Huang, G.-H., Chang, C.-T., Bilgihan, A., & Okumus, F. (2020). Helpful or harmful? A double-edged sword of emoticons in online review helpfulness. *Tourism Management, 81*, 104135.

Juuti, M., Sun, B., Mori, T., & Asokan, N. (2018). Stay on-topic: Generating context-specific fake restaurant reviews. In *European symposium on research in computer security (ESORICS)* (pp. 132–151).

Karakaya, F., & Barnes, N. (2010). Impact of online reviews of customer care experience on brand or company selection. *Journal of Consumer Marketing, 27*(5), 447–457.

Lerman, K., Yan, X., & Yu, X.-Z. (2016). The "majority illusion" in social networks. *PLoS One, 11*(2). https://doi.org/10.1371/journal.pone.0147617

Liu, Y., & Hu, H.-f. (2021). Online review helpfulness: The moderating effects of review comprehensiveness. *International Journal of Contemporary Hospitality Management, 33*(2), 534–556. https://doi.org/10.1108/IJCHM-08-2020-0856

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science, 62*(12). https://doi.org/10.1287/mnsc.2015.2304

Mayzlin, D. (2006). Promotional chat on the internet. *Marketing Science, 25*(2), 155–163.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review, 104*(8), 2421–2455.

O'Connor, P. (2008). User-generated content and travel: A case study on TripAdvisor.com. In P. O'Connor, W. Höpken, & U. Gretzel (Eds.), *Information and Communication Technologies in Tourism 2008* (pp. 47–58). Vienna: Springer.

Soto-Acosta, P. (2020). COVID-19 pandemic: Shifting digital transformation to a high-speed gear. *Information Systems Management, 37*(4). https://doi.org/10.1080/10580530.2020.1814461

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 40–53.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management, 58*, 51–65.

Yachin, J. (2018). The 'customer journey': Learning from customers in tourism experience encounters. *Tourism Management Perspectives, 28*, 201–210.

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management, 28*(1), 180–182.

Yoo, K. H., & Gretzel, U. (2009). Comparison of deceptive and truthful travel reviews. In W. Höpken, U. Gretzel, & R. Law (Eds.), *Information and communication technologies in tourism 2009*. Vienna: Springer.

Aarni Tuomi is a Lecturer at Haaga-Helia University of Applied Sciences, Finland. His research explores the intersection of emerging technology and human behaviour in service management and marketing contexts.