OXFORD

Interacting with Computers, 2023, **34**, 129–136
https://doi.org/10.1093/iwc/iwad010
Advance access publication date 18 February 2023
Article

Facilitating Implicit Emotion Regulation in Online News Commenting—An Experimental Vignette Study

Aleksi H. Syrjämäki [1,*], Mirja Ilves[1], Joel Kiskola [2], Anna Rantasila[2], Poika Isokoski[1], Thomas Olsson[2] and Veikko Surakka[1]

[1]Research Group for Emotions, Sociality, and Computing, Faculty of Information Technology and Communication Sciences, Tampere University, Finland
[2]Technology and Social Interaction Research Group, Faculty of Information Technology and Communication Sciences, Tampere University, Finland
*Corresponding author: aleksi.syrjamaki@tuni.fi

Abstract

An online experiment investigated the perceived effects of a user interface (UI) intervention aiming to support online news commenters' emotion regulation. By describing the comment's tone to the user, the expected effect was activation of the implicit emotion regulation process of affect labeling (i.e. naming emotions). The perceived emotion- and behavior-related effects of the labeling intervention were investigated using the experimental vignette methodology. Participants read a vignette describing the behavior of an uncivil commenter and assessed the commenter's probable responses to the labeling intervention or a control intervention shown in the UI. The results showed that, when compared to a control condition, the labeling intervention was assessed to evoke positive emotions and to result in mitigation of uncivil behavior. This suggests that UI solutions that support emotion regulation are a promising approach to reducing uncivil comments that users might afterward regret, and hence potentially improving the quality of online discussions.

RESEARCH HIGHLIGHTS:

- User interfaces that intervene in online commenting by describing online comments' tone could persuade users to reflect on their writing.
- In an experiment, an intervention was perceived as evoking positive emotions and mitigating incivility.
- User interfaces that support emotion regulation are a promising approach to improving online discussions.

Keywords: user interface, incivility, affect, computer-mediated communication, social media

1. INTRODUCTION

1.1. Background

Online news articles typically feature a comment section where users can discuss the article. It is widely known that many comments contain incivility, that is, norm-violating communication, such as disrespectful expressions toward other people or discussion topics (Coe *et al.*, 2014; Santana, 2019). Uncivil commenting can have various negative consequences, such as causing regret in the commenter (Wang *et al.*, 2011), evoking negative emotional experiences in the readers (Gervais, 2015), lowering the perceived quality of news sites (Prochazka *et al.*, 2018) and polarizing individuals' opinions on controversial subjects (Anderson *et al.*, 2014). Therefore, news sites have deployed various means against such behavior. Common solutions include the use of human moderators that delete inappropriate content (Boberg *et al.*, 2018) or a requirement to sign in to prevent anonymous commenting (Santana, 2019). While these solutions may be effective at reducing the overall number of uncivil comments, they have not fully solved the issue.

It should be noted that only some forms of incivility should be seen as problems that warrant intervention. In many cases, even highly uncivil user comments are perfectly legitimate and rational expressions of emotions, for instance due to experienced injustice (Bailey, 2018). Suppressing these types of expressions is harmful, particularly for disenfranchised people, who often experience 'tone policing' in online settings, wherein other users criticize the user's tone, rather than their argument (Davis and Ernst, 2019). Other forms of uncivil behavior can be more problematic. For instance, some people intentionally 'troll' to elicit reactions in other users and to cause disorder within discussions (Hardaker, 2010; Buckels *et al.*, 2014). In the current study, however, we are interested in another type of uncivil behavior, which we will discuss next.

Research on experiences of social media users suggest that people often write uncivil posts that they will later regret when they are in a 'hot', highly emotional state (Wang *et al.*, 2011). Experimental and quantitative evidence supports the notion that emotions are one of the psychological processes that cause uncivil commenting. Reading uncivil and emotionally negative comments evoke negative emotional reactions (Gervais, 2015; Masullo Chen and Ng, 2017) which, in turn, may motivate users to write further messages with a similar tone (Chmiel *et al.*, 2011; Kramer *et al.*, 2014; Cheng *et al.*, 2017; Ziegele *et al.*, 2018). Furthermore, strong emotional expressions cause online discussions to shift off-topic (Topal *et al.*, 2016). Seering *et al.* (2019) suggested that uncivil commenting could be reduced by methods that intervene with users' emotions. They developed

Received: March 8, 2022. **Revised:** September 15, 2022

user interface (UI) elements that evoke positive emotional responses in users (i.e. CAPTCHAs containing emotionally positive stimuli). In two experiments, it was found that participants exposed to these UI elements subsequently wrote more positive messages than participants exposed to no intervention. This result suggests that innovative UI intervention solutions can mitigate uncivil commenting by modulating user emotions.

Emotions are influenced by emotion-evoking stimuli, but perhaps more importantly, by various emotion regulation processes. Emotion regulation refers to psychological processes that alter the quality, intensity, or duration of emotions (e.g. Gross, 1998). Emotion regulation can be divided into explicit (intentional) and implicit (automatic) processes (Gyurak *et al.*, 2011). One of the implicit forms of emotion regulation is affect labeling, that is, naming of emotions (Torre and Lieberman, 2018). Psychological studies have shown that emotions are attenuated implicitly by giving a name to one's emotional state, or even by reading a label indicating the emotional characteristics of emotion-evoking stimuli (e.g. Hariri *et al.*, 2000; Lieberman *et al.*, 2007, 2011; Herbert *et al.*, 2013). Importantly for the present research, Fan *et al.* (2019) found that affect labeling modulates emotions in social media environments. In an analysis of the tone of Twitter posts, they found that after users had described their emotions (e.g. posted about feeling agitated), the users subsequently started posting more emotionally neutral content than before the naming of the emotion.

It has recently been proposed that news commenting platforms could utilize UIs that intentionally activate users' emotion regulation process of affect labeling (Kiskola *et al.*, 2021). This could tone down users' negative emotions and the resulting uncivil commenting behavior. Such systems could function as follows: the comment platform software analyzes the tone of user comments and then describes the emotional characteristics of the comment to the user. In line with the psychological affect labeling literature (Torre and Lieberman, 2018), it has been presumed that making the emotional elements of comments perceivable to users would help in emotion regulation (Kiskola *et al.*, 2021). A similar concept is used in Google's Perspective API, which automatically detects 'toxicity' in online comments (Jigsaw, 2021). By triggering notifications, it aims to persuade users to reflect on their writing. This type of intervention is potentially effective at mitigating the type of uncivil comments that people may afterward regret due to posting in a highly emotional state (see Wang *et al.*, 2011). However, we are not aware of any published research investigating how UIs that facilitate affect labeling would influence users' emotions and behavior. The current study addresses this gap with a vignette experiment.

1.2. Present study

1.2.1. *UI interventions*

This study examined users' perceptions of the emotional and behavioral effects of a potential UI intervention utilizing affect labeling. The labeling intervention of this study notifies an uncivil commenter that their comment had been analyzed and describes the tone of the comment to them (a more detailed description will be provided in Section 2.2.). The labeling intervention aims to facilitate emotion regulation and to provide the user with an opportunity to consider whether they want to post the comment. The intention is to mitigate the types of comments that people may regret afterward due to posting in a heated, emotional state (Wang *et al.*, 2011).

The effects of the labeling intervention were compared to two control interventions. 1) The guidelines-control intervention was a simple reference to posting guidelines. This allows determining whether the effects of the labeling intervention differ from effects of a typical UI element that supposedly aims to reduce uncivil commenting. 2) The analysis-control intervention notifies an uncivil user that their comment had been analyzed, and then reminds the user of the posting guidelines. It was essential to include a control condition that involves an analysis of the comment to allow disentangling effects evoked by affect labeling from effects evoked by the analysis. This is because the analysis of a comment could evoke an experience of being observed which could, by itself, reduce inappropriate behavior (Bateson *et al.*, 2006; Nettle *et al.*, 2012; Priks, 2014; Jansen *et al.*, 2018).

While the current study investigates perceived reactions to potential UI elements, the presented UI interventions are not meant as contributions to UI design as such. Rather, they were created for the purposes of the current experiment. We primarily considered the comparability of the experimental conditions, rather than any design principles, such as usability or aesthetics.

1.2.2. *Experimental vignette methodology*

To compare the perceived effects of the interventions, the present study utilized the experimental vignette methodology (Atzmüller and Steiner, 2010; Aguinis and Bradley, 2014). In this methodology, participants are presented with vignettes, which are carefully constructed short descriptions or 'stories' of hypothetical scenarios. Participants then answer questions regarding the vignette, indicating for instance what they think about the scenario, or evaluating how people would act in the situation (Hughes, 1998; Aguinis and Bradley, 2014). While the vignettes are fictional, they are useful for gauging people's perceptions, beliefs and attitudes about various situations and phenomena (Schoenberger and Ravdal, 2000; Atzmüller and Steiner, 2010). Critically, the vignettes can be experimentally manipulated so that participants are shown slightly different versions of the vignettes in different experimental conditions. This allows investigation of causal relationships, providing evidence of what shapes participants' perceptions and beliefs about the scenario.

Vignette methodology is widely used in various disciplines to investigate people's perceptions of diverse social issues and problems (Barter and Renold, 2000; Schoenberger and Ravdal, 2000). In addition to studying people's intentions, attitudes and behaviors, experimental vignettes have been used to study emotions, for example, emotion understanding (Schlegel & Scherer, 2017) and psychophysiological emotional responses (Krumhuber *et al.*, 2018). Recently, Wingenbach *et al.* (2019) created and validated verbal, text-based emotion vignettes to elicit specific emotions. They showed that the individual emotion vignettes clearly mapped onto distinct emotion categories and intensity rates for the self-reported emotional experience were high while participants immersed themselves in the scenarios depicted in the vignettes. Thus, vignettes seem to be appropriate method to elicit and study internal emotional responses.

An important strength of this methodology is that it can be used when measurement of real behavior would be difficult or unethical (Aguinis and Bradley, 2014). For the present study, it provided a way to study the probable emotional and behavioral effects of a UI intervention before developing a functioning system that analyzes user comments. Furthermore, the methodology allowed us to avoid the potential ethical challenges related to use of emotion manipulations in experimental research in real social media environments (Shaw, 2016). Notably, the methodology has been criticized for relatively low external validity, i.e. that evaluation of fictional scenarios may not reflect responses to comparable real-life situations (Aguinis and Bradley, 2014).

However, there is also evidence that vignette experiments can produce knowledge applicable to the real world (Hainmueller *et al.*, 2015). The study investigated behavior in referendums, where people can vote on whether a person should receive Swiss citizenship. It was found that voting behavior was quite similar in the real situation as when similar choices were presented in artificial vignettes. This finding suggests that at least in some contexts, responses to vignettes resemble responses in comparable real-life situations.

In this study, rather than assessing their own reactions to the UI interventions, the participants assessed the probable responses of a hypothetical uncivil user. This decision was made for the following reasons. First, people tend to believe that they themselves would show milder emotional responses to uncivil content than other people would (Masullo Chen and Ng, 2017). Second, people often give socially desirable responses in studies and might therefore be reluctant to admit behaving in an uncivil way themselves (see literature on social desirability bias, e.g, Nederhof, 1985). Therefore, if the participants assessed their own responses to the interventions, they might report themselves as being overly calm and rational.

1.2.3. *Study design and hypotheses*

The participants were presented with a vignette describing the behavior of an online user that ends up writing an uncivil comment in a news site's comment section. The vignette then described how, just before posting the comment, the user notices either the labeling, or the guidelines-control or analysis-control UI intervention. Following the vignette, participants filled in a questionnaire to assess how the user would respond to the intervention at the emotional and behavioral level. The impact on emotions was assessed in terms of change in the user's emotional state (toward a more positive/negative or aroused/unaroused state, in line with the dimensional emotion framework by Bradley and Lang, 1994, 2007). As for the impact on behavior, participants assessed what the user would do next, using a multiple-choice question. Some response options indicated that the intervention successfully mitigated uncivil behavior (e.g. the user would refrain from posting the uncivil comment), and others indicated that the intervention was unsuccessful (e.g. the user would post the uncivil comment).

We expected that when compared to the two control conditions, the labeling condition would result in a higher frequency of responses predicting successful mitigation of uncivil behavior. We also expected the analysis-control to be associated with a higher frequency of successes than the guidelines-control condition. Importantly, we expected that the labeling condition would be assessed to make the user feel more positive than the two control conditions. We also expected the emotion-related effect to mediate the behavioral effect of the labeling condition, which would be consistent with the notion that the intervention functions by regulating negative emotions.

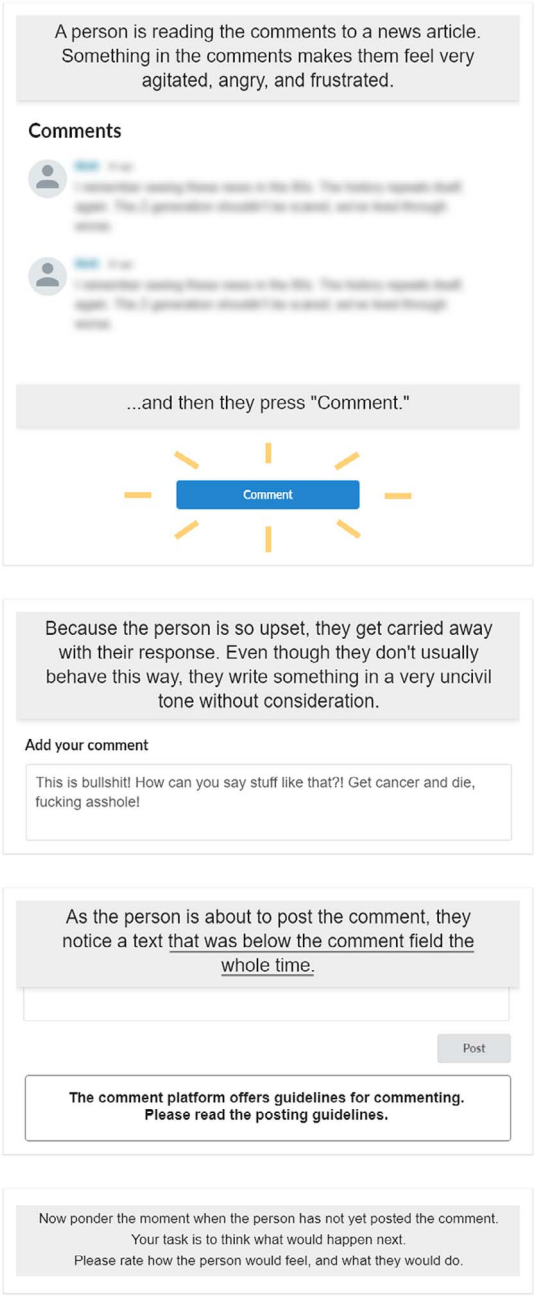

FIGURE 1. The vignette, as depicted in the guidelines-control group.

2. MATERIALS AND METHODS

2.1. Participants

We recruited 1411 participants via Prolific (www.prolific.co), with the following eligibility criteria: fluent in English, normal or corrected-to-normal vision and a minimum of 70% approval rate on Prolific. Participants were required to use desktop or laptop computers. They were rewarded with £0.63. Participants were randomly allocated to three experimental groups (labeling/analysis-control/guidelines-control; see Section 2.2. for more information about the experimental manipulations). Fifty-four participants failed an attention check, and an additional 324 participants failed one of the manipulation checks (see 2.3. for more information) and were removed from all analyses. Thus, the analyzed sample consisted of 1033 participants (M_{age} = 27.0, SD_{age} = 9.5, 609 males, 416 females, 8 other, $n_{labeling}$ = 386, $n_{analysis\text{-}control}$ = 357, $n_{guidelines\text{-}control}$ = 290).

2.2. Procedure

An ethical statement for the study procedures was obtained from the Ethics Committee of the Tampere Region. The experiment was run online using the LimeSurvey survey tool. After reading general information about the study, participants indicated informed consent by checking a box.

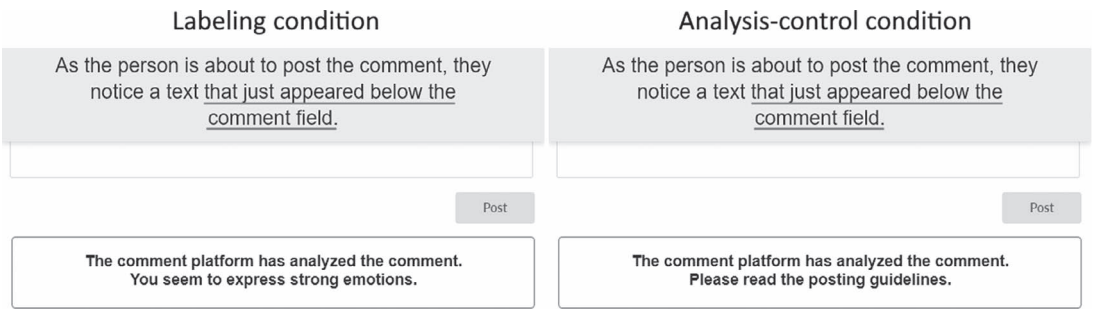

FIGURE 2. The third panel of the vignette, as depicted in labeling and analysis-control groups.

Participants were then presented with a vignette, describing a user reading online news comments. The vignette consisted of four panels, aligned on top of each other on the survey page. See Figure 1 for an illustration of the vignette, as depicted in the guidelines-control group. The first panel of the vignette showed the user reading news comments and feeling agitated in response. In the second panel, the user was described writing a highly uncivil comment (randomly one of the following: "This is bullshit! How can you say stuff like that?! Get cancer and die, fucking asshole!", "Fuck this shit! That's completely false! Burn in hell, you goddamn prick!", or "That's wrong! You're a fucking liar. Go hang yourself!").

The experimental manipulation was administered on the third panel of the vignette. The panel described how, just before posting the comment, the user noticed a text depicted in the UI of the commenting platform. In the labeling group, the text was described as having appeared as a response to the comment platform analyzing the content of the user's comment, and it indicated that the user had expressed strong emotions. In the analysis-control group, the text was similarly described as having appeared as a result of the analysis of the comment, but instead of describing the tone of the comment, it referred to posting guidelines. In the guidelines-control group, the text similarly referred to posting guidelines, but was described as having been visible below the comment field the whole time. See Figure 2 for the third panel of the vignette, as depicted in the labeling and analysis-control groups. In the final panel, the participant was instructed to consider what the user would next feel and do.

2.3. Measurements

After reading the vignette, participants filled in a few questionnaire items to assess the user's probable emotional and behavioral responses. In four items, the participants were to rate whether the user's emotional state would have changed to a more positive, negative, aroused and unaroused direction, when compared to how they felt while writing the uncivil comment. The rating scale was a seven-point scale varying from 1 (completely disagree) to 7 (completely agree). An attention check item was also included, in which participants were explicitly told to respond, 'completely agree'. As mentioned above, the participants who failed the attention check were excluded from all analyses. Next, the participants were asked in a multiple-choice item whether they believed the user would (1) send the comment without changes, (2) make the tone of the comment more positive and then send the comment, (3) make the tone of the comment more negative and then send the comment or (4) not send the comment. See the Appendix for a detailed description of the emotion and behavior measurements.

A manipulation check was administered on the following page. Participants were asked two questions about what happened in the vignette, with two response options for each. First, they were asked whether the text that the user noticed appeared after writing the comment, or whether it was visible the whole time. Second, they were asked whether the text referred to posting guidelines or the user's emotional expressions. As mentioned above, participants who failed either of the manipulation checks were excluded from all analyses. Finally, participants provided background information (age, gender) and were then thanked for participation.

2.4. Data analysis

To compare the effects of the interventions on emotion ratings, the ratings were analyzed with one-way ANOVAs with experimental condition as the independent variable. Significant effects (at $\alpha = .05$) were followed with Bonferroni corrected post-hoc pairwise comparisons using *t* tests.

The effects of the interventions on predicted behavior were analyzed as follows. Responses to the multiple-choice question were first recoded into a binary variable. Responses predicting that the user would change the tone of the comment more positive, or refrain from posting were coded as a success in mitigation of incivility. Responses predicting that the user would change the tone of the comment more negative, or post the comment as such, were coded as a failure in mitigation of incivility. To compare the rate of predicted successes and failures across experimental conditions, a 2 (success/failure) × 3 (labeling/analysis-control/guidelines-control) χ^2 test of homogeneity was conducted. For post-hoc pairwise comparisons between interventions, we conducted all three possible 2 × 2 χ^2 tests with Bonferroni correction to adjust for Type I error rate (MacDonald and Gardner, 2000; Sharpe, 2015).

When an intervention was found to have effects on both emotions and behavior as compared to the guidelines-control condition, we examined whether the behavioral effect was mediated by the emotional effect. We used the method by Iacobucci (2012) which allows investigation of mediation when categorical variables are included in the model. First, the strength of the direct path from the independent variable (i.e. the intervention of interest versus guidelines-control intervention as a dummy-coded variable) to the dependent variable (i.e. the predicted behavioral impact, success/failure) was estimated with a binary logistic regression model. After this, a linear regression model was fitted to determine the strength of the path from the independent variable to the mediator (the emotion rating). Then, another binary logistic regression model was fitted, with both the independent variable and the mediator as predictors. Finally, we computed the z-test and compared $z_{mediation}$ with the critical value of |1.96| to determine whether the mediation effect was significant at $\alpha = .05$ (Iacobucci, 2012).

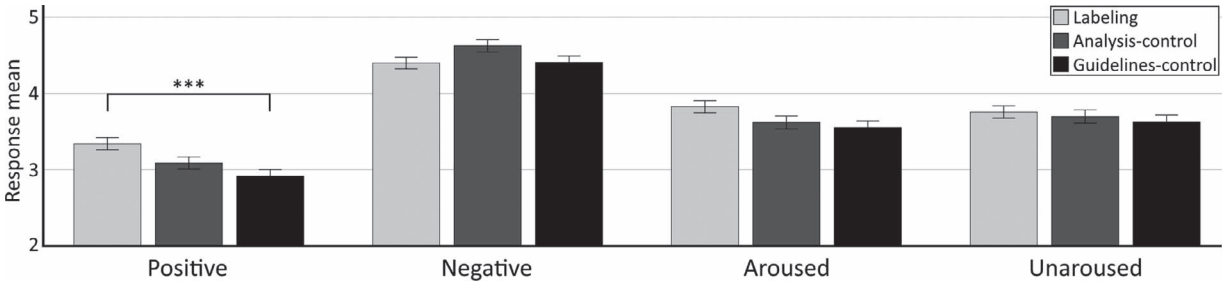

FIGURE 3. Means of emotion ratings in each condition. For each item, participants were indicating whether the user's emotions would have changed into the respective direction, when compared to how they felt while writing the uncivil comment. Responses were given on a 1 (completely disagree) to 7 (completely agree) scale. The error bars denote standard error of the means. Significant differences between groups are pointed out. ***p < .001.

TABLE 1. Frequencies of predicted successes and failures in each experimental group. The left side of the table shows predicted successes, both aggregated into a single variable (as used in the analyses), and the two response options separately. The predicted failures are similarly shown on the right side of the table

	Success	Make more positive and send	Not send	Failure	Make more negative and send	Send without changes
Labeling	237	142	95	149	32	117
Analysis-control	244	125	119	113	19	94
Guidelines-control	148	81	67	142	3	139

TABLE 2. Results of pairwise comparisons comparing effects of the interventions on mitigation of uncivil behavior

Comparison	$\chi^2(1)$	p
Labeling-guidelines	7.3	.007
Labeling-analysis	3.9	.048
Analysis-guidelines	20.1	< .001

Note. Bonferroni-adjusted α = .017

3. RESULTS

3.1. Perceived emotional impact

See Figure 3 for means of the emotion-related ratings in each condition. One-way ANOVAs revealed a significant effect of Intervention on positive emotions ($F(2, 1030) = 6.71$, $p = .001$, $\eta^2 = .013$), but not on any of the other emotion-related ratings (lowest p was for the effect on arousal rating, $F(2, 1030) = 2.94$, $p = .053$, $\eta^2 = .006$; other ps > 0.07). Pairwise comparisons showed that the ratings of positive emotions were higher in the labeling group than in the guidelines-control group ($t(674) = 3.62$, $p < 0.001$, $d = 0.28$, $\alpha_{adjusted} = .017$), but no other significant differences were found (lowest p was for the comparison between labeling and analysis-control; $t(741) = 2.23$, $p = 0.026$, $d = 0.16$, $\alpha_{adjusted} = .017$).

3.2. Perceived behavioral impact

For frequencies of predicted successes and failures at mitigation of incivility in each experimental condition, see Table 1. A χ^2 test showed that there were significant differences between conditions in the frequency of successes and failures ($\chi^2(2) = 20.2$, $p < .001$). Post-hoc comparisons showed that both labeling and analysis-control conditions were associated with a higher frequency of successes vs. failures, when compared to the guidelines-control condition, but the two interventions did not differ significantly from each other (see Table 2).

3.3. Mediation analysis

Figure 4 depicts the models describing the effect of the labeling intervention on behavior as a direct effect (top) and as an effect mediated by positive emotion (bottom). Most importantly, a z-test showed that the mediation effect was statistically significant at $\alpha = .05$ ($z_{mediation} = 3.33 > |1.96|$). Interestingly, the direct effect from the labeling intervention to behavior was not statistically significant when the mediator was included in the model. This suggests that the behavior-related effect of the intervention was fully mediated by its perceived effect on positive emotion (indirect-only mediation; Zhao *et al.*, 2010).

4. DISCUSSION

As hypothesized, both the labeling intervention and the analysis-control intervention were assessed to mitigate uncivil commenting more than the guidelines-control intervention. The finding is in line with a previous research that has shown that these kinds of small-scale interventions have potential to mitigate the negative effects caused by incivility. For example, Yeo *et al.* (2019) found that only cues about comment moderation without any actual change in the comments themselves alleviated the effect of uncivil comments on perceptions of news bias.

Contrary to our expectations the labeling intervention did not significantly differ from the analysis-control intervention in terms of perceived behavioral effects. Importantly, however, the results suggest that the two interventions were seen as operating via different psychological mechanisms. The labeling intervention was assessed to evoke more positive emotions than the guidelines-control intervention, and the emotional effect was mediating the behavior-related effect. In contrast, the analysis-control intervention did not influence emotion ratings when compared to the guidelines-control condition. This shows that the behavior-related effect of the labeling intervention was driven

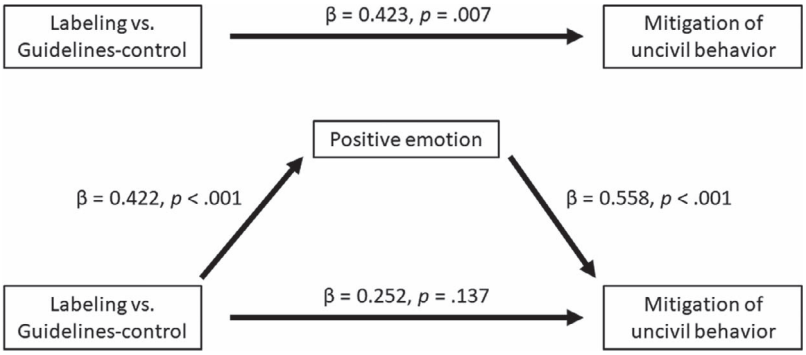

FIGURE 4. Models describing a direct effect of the labeling intervention on predicted behavior (top), and the effect mediated by positive emotion (bottom).

by its perceived effect on emotions, rather than it reflecting the effect of the analysis of the comment as such, which could have mitigated uncivil behavior by making the user feel being observed (Bateson *et al.*, 2006; Jansen *et al.*, 2018).

Prior research has shown that when social media users label emotions, their emotional state is attenuated (Fan *et al.*, 2019). The current result suggests that affect labeling can also be facilitated with UI interventions that describe the tone of a comment to the user (as also suggested by Jigsaw, 2021; Kiskola *et al.*, 2021). This is perceived as effective at attenuating negative emotions and motivating civil commenting. Perspective API is one example of solutions based on computational approach, which aim to regulate the tone of discussion. It can detect and inform user about 'toxic' writing, and it has been reported to have some positive impact on the quality of discussion (Goldberg *et al.*, 2020). Further, Peng *et al.* (2019) designed the GremoBot chatbot to support emotion regulation in group chats. The results of their study suggest that visualizing group emotion to the participants can be useful for enhancing positive feelings and steering them away from negative words. Together with these findings, our results suggest that interventions facilitating emotion regulation are a promising way to complement other approaches to improving the quality of news commenting. It is noteworthy that affect labeling mitigates emotions without requiring any explicit effort from the users (Torre and Lieberman, 2018). Importantly, this makes the intervention relatively unintrusive from the user's point of view. Automatic systems based on affect labeling may also benefit moderators by reducing their workload. For news organizations, an intervention that does not restrict users' ability to post what they want may help maintain an image of neutrality (Wolfgang *et al.*, 2020).

In the present study, we investigated very simple, text-based interventions. It was important to strip down the interventions to their essentials to keep factors, such as the visual appearance of the interventions similar across experimental conditions. This makes it possible to conclude that the perceived effects of the labeling intervention were driven by the description of the tone of the comment, rather than by some other difference between the conditions, such as the visual appearance of the UI. Of course, this type of intervention mechanism could be utilized in various ways in UI design. For instance, UI designers could vary the contents of the notification (e.g. informing the user of how others would feel about the post, instead of only describing its tone), its visual appearance (e.g. using pictures or symbols instead of text), or timing (e.g. during writing of the comment, rather than after it). Kiskola *et al.* (2021) provided a more extensive exploration of the potential design space in this context. Future research could examine how perceptions of emotion- and behavior-related effects of affect labeling interventions are influenced by these different design choices. This could help further refine the efficacy of the intervention.

To reflect on the methodology, the experimental vignette methodology provided a cost-effective and ethical way to investigate the perceived emotional and behavioral effects of the labeling intervention in the present study. The methodology is useful for drawing causal inferences about effects of specific variables on participants' perceptions of hypothetical scenarios (Atzmüller and Steiner, 2010). However, further research is needed to determine how the intervention would influence actual user behavior in a real social media environment.

Before deploying labeling interventions on real social media services, it would be crucial to investigate how the intervention influences different kinds of users. In the current study, participants assessed how the interventions would influence a user who wrote a highly uncivil comment in response to a negative emotional reaction. However, this represents only one type of commenter. Sometimes uncivil comments can also be perfectly rational reactions to experienced injustice (Bailey, 2018). It would be important to ensure that the intervention is not experienced as condescending 'tone policing' in such cases. Furthermore, future research should investigate how the intervention would influence more civil commenters, or commenters who intentionally 'troll' others for their own amusement (see Buckels *et al.*, 2014).

5. CONCLUSION

Improving the quality of online news commenting is important for news organizations, who can genuinely suffer from the prevalence of uncivil and hostile comments (Prochazka *et al.*, 2018). In the current study, we demonstrated that the presented UI facilitating users' implicit emotion regulation through affect labeling was perceived effective in motivating civil commenting behavior. Such UI solutions appear a promising approach to nurturing constructive and civil discussions on news commenting platforms. Civility in discussions can have various positive effects, such as increasing users' willingness to participate in the discussions (Molina and Jennings, 2018) and to share additional perspectives on the topic (Han *et al.*, 2018). Improvements in the civility of discussions can thus help develop online discussions toward a medium where everyone has a voice. This may benefit the whole society, making it more inclusive and safer for everyone.

Funding

This work was supported by the Academy of Finland under grants 320766 and 320767.

Conflict of interest

The authors report no conflict of interest.

Supplementary material

Supplementary data is available at *Interacting with Computers* online.

References

Aguinis, H. and Bradley, K. J. (2014) Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organ. Res. Methods*, **17**, 351–371. https://doi.org/10.1177/1094428114547952.

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A. and Ladwig, P. (2014) The "nasty effect:" online incivility and risk perceptions of emerging technologies. *J. Comput.-Mediat. Commun.*, **19**, 373–387. https://doi.org/10.1111/jcc4.12009.

Atzmüller, C. and Steiner, P. M. (2010) Experimental vignette studies in survey research. *Methodology*, **6**, 128–138. https://doi.org/10.1027/1614-2241/a000014.

Bailey, A. (2018) On anger, silence, and epistemic injustice. *R. Inst. Philos. Suppl.*, **84**, 93–115. https://doi.org/10.1017/S1358246118000565.

Barter, C. and Renold, E. (2000) 'I wanna tell you a story': exploring the application of vignettes in qualitative research with children and young people. *Int. J. Soc. Res. Methodol.*, **3**, 307–323. https://doi.org/10.1080/13645570050178594.

Bateson, M., Nettle, D. and Roberts, G. (2006) Cues of being watched enhance cooperation in a real-world setting. *Biol. Lett.*, **2**, 412–414. https://doi.org/10.1098/rsbl.2006.0509.

Boberg, S., Schatto-Eckrodt, T., Frischlich, L. and Quandt, T. (2018) The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media Commun.*, **6**, 58–69. https://doi.org/10.17645/mac.v6i4.1493.

Bradley, M. M. and Lang, P. J. (1994) Measuring emotion: the self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry*, **25**, 49–59. https://doi.org/10.1016/0005-7916(94)90063-9.

Bradley, M. M., & Lang, P. J. (2007). Emotion and motivation. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 581–607). Cambridge University Press. https://doi.org/10.1017/CBO9780511546396.025

Buckels, E. E., Trapnell, P. D. and Paulhus, D. L. (2014) Trolls just want to have fun. *Personal. Individ. Differ.*, **67**, 97–102. https://doi.org/10.1016/j.paid.2014.01.016.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1217-1230). https://doi.org/10.1145/2998181.2998213

Chmiel, A., Sienkiewicz, J., Thelwall, M., Paltoglou, G., Buckley, K., Kappas, A. and Hołyst, J. A. (2011) Collective emotions online and their influence on community life. *PLoS One*, **6**, e22207. https://doi.org/10.1371/journal.pone.0022207.

Coe, K., Kenski, K. and Rains, S. A. (2014) Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *J. Commun.*, **64**, 658–679. https://doi.org/10.1111/jcom.12104.

Davis, A. M. and Ernst, R. (2019) Racial gaslighting. *Politics, Groups, and Identities*, **7**, 761–774. https://doi.org/10.1080/21565503.2017.1403934.

Fan, R., Varol, O., Varamesh, A., Barron, A., van de Leemput, I. A., Scheffer, M. and Bollen, J. (2019) The minute-scale dynamics of online emotions reveal the effects of affect labeling. *Nat. Hum. Behav.*, **3**, 92–100. https://doi.org/10.1038/s41562-018-0490-5.

Gervais, B. T. (2015) Incivility online: affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics*, **12**, 167–185. https://doi.org/10.1080/19331681.2014.997416.

Goldberg, I., Simon, G. and Thimmaiah, K. (2020) OpenWeb tests the impact of "nudges" in online discussions. Retrieved January 12, 2023, from OpenWeb website: https://www.openweb.com/blog/openweb-improves-community-health-with-real-time-feedback-powered-by-jigsaws-perspective-api/.

Gross, J. J. (1998) The emerging field of emotion regulation: an integrative review. *Rev. Gen. Psychol.*, **2**, 271–299. https://doi.org/10.1037/1089-2680.2.3.271.

Gyurak, A., Gross, J. J. and Etkin, A. (2011) Explicit and implicit emotion regulation: a dual-process framework. *Cognit. Emot.*, **25**, 400–412. https://doi.org/10.1080/02699931.2010.544160.

Hainmueller, J., Hangartner, D. and Yamamoto, T. (2015) Validating vignette and conjoint survey experiments against real-world behavior. *Proc. Natl. Acad. Sci.*, **112**, 2395–2400. https://doi.org/10.1073/pnas.1416587112.

Han, S. H., Brazeal, L. M. and Pennington, N. (2018) Is civility contagious? Examining the impact of modeling in online political discussions. *Social Media+ Society*, **4**, 2056305118793404. https://doi.org/10.1177/2056305118793404.

Hardaker, C. (2010) Trolling in asynchronous computer-mediated communication: from user discussions to academic definitions. *Journal of Politeness Research*, **6**, 215–242. https://doi.org/10.1515/JPLR.2010.011.

Hariri, A. R., Bookheimer, S. Y. and Mazziotta, J. C. (2000) Modulating emotional responses: effects of a neocortical network on the limbic system. *Neuroreport*, **11**, 43–48.

Herbert, C., Sfärlea, A. and Blumenthal, T. (2013) Your emotion or mine: Labeling feelings alters emotional face perception—an ERP study on automatic and intentional affect labeling. *Front. Hum. Neurosci.*, **7**, 378. https://doi.org/10.3389/fnhum.2013.00378.

Hughes, R. (1998) Considering the vignette technique and its application to a study of drug injecting and HIV risk and safer behaviour. *Sociology of health & illness*, **20**, 381–400. https://doi.org/10.1111/1467-9566.00107.

Iacobucci, D. (2012) Mediation analysis and categorical variables: the final frontier. *J. Consum. Psychol.*, **22**, 582–594. https://doi.org/10.1016/j.jcps.2012.03.006.

Jansen, A. M., Giebels, E., van Rompay, T. J. and Junger, M. (2018) The influence of the presentation of camera surveillance on cheating and pro-social behavior. *Front. Psychol.*, **9**, 1937. https://doi.org/10.3389/fpsyg.2018.01937.

Jigsaw (2021) Perspective API. Retrieved January 12, 2023, from https://www.perspectiveapi.com/#/home.

Kiskola, J., Olsson, T., Väätäjä, H., Syrjämäki, A., Rantasila, A., Isokoski, P., Ilves, M. and Surakka, V. (2021) Applying critical voice in design of user interfaces for supporting self-reflection and emotion regulation in online news commenting. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445783.

Kramer, A. D., Guillory, J. E. and Hancock, J. T. (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 8788–8790. https://doi.org/10.1073/pnas.1320040111.

Krumhuber, E. G., Tsankova, E. and Kappas, A. (2018) Examining subjective and physiological responses to norm violation using text-based vignettes. *Int. J. Psychol.*, **53**, 23–30. https://doi.org/10.1002/ijop.12253.

Lieberman, M. D., Eisenberger, N. I., Crockett, M. J., Tom, S. M., Pfeifer, J. H. and Way, B. M. (2007) Putting feelings into words. *Psychol. Sci.*, **18**, 421–428. https://doi.org/10.1111/j.1467-9280.2007.01916.x.

Lieberman, M. D., Inagaki, T. K., Tabibnia, G. and Crockett, M. J. (2011) Subjective responses to emotional stimuli during labeling, reappraisal, and distraction. *Emotion*, **11**, 468–480. https://doi.org/10.1037/a0023503.

MacDonald, P. L. and Gardner, R. C. (2000) Type I error rate comparisons of post hoc procedures for I J chi-square tables. *Educ. Psychol. Meas.*, **60**, 735–754. https://doi.org/10.1177/00131640021970871.

Masullo Chen, G. and Ng, Y. M. M. (2017) Nasty online comments anger you more than me, but nice ones make me as happy as you. *Comput. Hum. Behav.*, **71**, 181–188. https://doi.org/10.1016/j.chb.2017.02.010.

Molina, R. G. and Jennings, F. J. (2018) The role of civility and metacommunication in Facebook discussions. *Commun. Stud.*, **69**, 42–66. https://doi.org/10.1080/10510974.2017.1397038.

Nederhof, A. J. (1985) Methods of coping with social desirability bias: a review. *Eur. J. Soc. Psychol.*, **15**, 263–280. https://doi.org/10.1002/ejsp.2420150303.

Nettle, D., Nott, K. and Bateson, M. (2012) 'Cycle thieves, we are watching you': impact of a simple signage intervention against bicycle theft. *PLoS One*, **7**, e51738. https://doi.org/10.1371/journal.pone.0051738.

Peng, Z., Kim, T. and Ma, X. (2019) GremoBot: exploring emotion regulation in group chat. *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, 335–340. https://doi.org/10.1145/3311957.3359472.

Priks, M. (2014) Do surveillance cameras affect unruly behavior? A close look at grandstands. *Scand. J. Econ.*, **116**, 1160–1179. https://doi.org/10.1111/sjoe.12075.

Prochazka, F., Weber, P. and Schweiger, W. (2018) Effects of civility and reasoning in user comments on perceived journalistic quality. *Journal. Stud.*, **19**, 62–78. https://doi.org/10.1080/1461670X.2016.1161497.

Santana, A. D. (2019) Toward quality discourse: measuring the effect of user identity in commenting forums. *Newsp. Res. J.*, **40**, 467–486. https://doi.org/10.1177/0739532919873089.

Schlegel, K. and Scherer, K. R. (2018) The nomological network of emotion knowledge and emotion understanding in adults: evidence from two new performance-based tests. *Cognit. Emot.*, **32**, 1514–1530. https://doi.org/10.1080/02699931.2017.1414687.

Schoenberger, N. E. and Ravdal, H. (2000) Using vignettes in awareness and attitudinal research. *Int. J. Soc. Res. Methodol.*, **3**, 63–74. https://doi.org/10.1080/136455700294932.

Seering, J., Fang, T., Damasco, L., Chen, M., Sun, L., & Kaufman, G. (2019). Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-14). https://doi.org/10.1145/3290605.3300836

Sharpe, D. (2015) Chi-square test is statistically significant: now what? *Pract. Assess. Res. Eval.*, **20**, 8. https://doi.org/10.7275/tbfa-x148.

Shaw, D. (2016) Facebook's flawed emotion experiment: antisocial research on social network users. *Research Ethics*, **12**, 29–34. https://doi.org/10.1177/1747016115579535.

Topal, K., Koyuturk, M., & Ozsoyoglu, G. (2016). Emotion-and area-driven topic shift analysis in social media discussions. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 510–518). IEEE. https://doi.org/10.1109/ASONAM.2016.7752283

Torre, J. B. and Lieberman, M. D. (2018) Putting feelings into words: affect labeling as implicit emotion regulation. *Emot. Rev.*, **10**, 116–124. https://doi.org/10.1177/1754073917742706.

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G. and Cranor, L. F. (2011) " I regretted the minute I pressed share" a qualitative study of regrets on Facebook. *Proceedings of the seventh symposium on usable privacy and security*, 1–16. https://doi.org/10.1145/2078827.2078841.

Wingenbach, T. S., Morello, L. Y., Hack, A. L. and Boggio, P. S. (2019) Development and validation of verbal emotion vignettes in Portuguese, English, and German. *Front. Psychol.*, **10**, 1135. https://doi.org/10.3389/fpsyg.2019.01135.

Wolfgang, J. D., Blackburn, H. and McConnell, S. (2020) Keepers of the comments: how comment moderators handle audience contributions. *Newsp. Res. J.*, **41**, 433–454. https://doi.org/10.1177/0739532920968338.

Yeo, S. K., Su, L. Y. F., Scheufele, D. A., Brossard, D., Xenos, M. A. and Corley, E. A. (2019) The effect of comment moderation on perceived bias in science news. *Inf. Commun. Soc.*, **22**, 129–146. https://doi.org/10.1080/1369118X.2017.1356861.

Zhao, X., Lynch, J. G., Jr. and Chen, Q. (2010) Reconsidering baron and Kenny: myths and truths about mediation analysis. *J. Consum. Res.*, **37**, 197–206. https://doi.org/10.1086/651257.

Ziegele, M., Weber, M., Quiring, O. and Breiner, T. (2018) The dynamics of online news discussions: effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions. *Inf. Commun. Soc.*, **21**, 1419–1435. https://doi.org/10.1080/1369118X.2017.1324505.