The Seventh AAAI Conference on Human
Computation and Crowdsourcing (HCOMP-19)

Testing Stylistic Interventions to Reduce Emotional Impact of Content Moderation Workers

Sowmya Karunakaran, Rashmi Ramakrishan
Google Inc
{sowmyakaru, raramakrishnan}@google.com

Abstract

With the rise in user generated content, there is a greater need for content reviews. While machines and technology play a critical role in content moderation, the need for manual reviews still remains. It is known that such manual reviews could be emotionally challenging. We test the effects of simple interventions like grayscaling and blurring to reduce the emotional impact of such reviews. We demonstrate this by bringing in interventions in a live content review setup thus allowing us to maximize external validity. We use a pre-test post-test experiment design and measure review quality, average handling time and emotional affect using the PANAS scale. We find that simple grayscale transformations can provide an easy to implement and use solution that can significantly change the emotional impact of content reviews. We observe, however, that a full blur intervention can be challenging to reviewers.

Introduction

With the rise in user generated content, there has been a significant increase in content shared online every day through social networks and content platforms. This in turn has increased the need to moderate content to ensure it complies with community guidelines and policies. Content moderation relies on automated processes and manual reviews by human reviewers to determine if content displayed in the form of images, videos or text, violate the platform's acceptable use policies. For example, on Google Drive, Photos and Blogger, in the past year, 160,000 pieces of violent extremism content were taken down (Canegallo 2019). At the extreme, such content can include ultra-graphic violent acts, such as murder, suicide, and animal abuse (Chen 2014; Krause and Grassegger 2016).

While machines and technology play a critical role in content moderation, there continues to be a need for manual reviews where human judgment is required in interpreting borderline cases as well as generation of ground truth for machine learning models. It is known, however, that such manual reviews could be emotionally challenging. While a large chunk of content posted on online platforms is safe, certain pieces of content could be violating the platforms acceptable use policies and community guidelines.

There is an abundance of research on automated ways to detect and filter out such unwanted or harmful content. Many such applications use algorithms that have been built using machine learning approaches. While a lot of progress has been made over the last decade, there is still a need for human involvement. To more reliably moderate user content, social media companies hire internal reviewers, contract specialized workers from third parties, or outsource to online labor markets (Gillespie 2017; Roberts 2016). Further, the subjectivity and ambiguity of the moderation tasks make it inevitable to use manual reviewers as opposed to a pure algorithmic review system (Roberts 2017; 2018). Many such manual review processes involve reviewing difficult and challenging content. It is known that extensive viewing of such disturbing content can incur significant health consequences for the reviewers. We discuss this further in the related work section.

Despite the importance of the subject, there is no prior research on the effects of technical interventions to reduce the associated emotional impact to reviewers. To address this gap, we present a study measuring the emotional impact of reviewing difficult content by introducing simple image transformations such as grayscaling and blurring of content. Figure 1 shows a sample of grayscaling and blurring transformations. We conduct a series of experiments on live content review queues. We maximize external validity by studying the impact on live manual review queues and test for differences in emotions, output quality, and task completion times with respect to our interventions. Our main results are the following:

- Grayscaling of images under review significantly improves positive affect for reviewers without leading to significant decline in business metrics.
- Blurring of images under review depleted the positive emotional affect for reviewers and increased irritability.
- We find that the PANAS scale serves as a useful self-reporting mechanism to be applied in the context of measuring changes in emotional affect as a result of engaging in difficult content moderation tasks.

Figure 1: Sample Image showing the effects of the treatments: a) Original image b) Grayscale and c) Blurred.

Related Work

Manual Reviews and Challenges

Content moderation work involves reviewing for various types of badness or policy violations and unacceptable use of the online platform. Such moderation tasks often involve reviewing difficult content. For example, hate speech detection and text civility is a common moderation task for humans and machines (Rojas-Galeano 2017; Schmidt and Wiegand 2017). Violence detection in images and videos is another common moderation task for humans and machines (Deniz et al. 2014; Gao et al. 2016). This work is expected to be generally unpleasant. There is, however, an increasing awareness and recognition that beyond mere unpleasantness, long-term or extensive viewing of such disturbing content can incur significant health consequences for those engaged in such tasks (Chen 2014; Ghoshal 2017). Such incidental emotions have been shown to even influence everyday activities such as how people make decisions (Vohs, Baumeister, and Loewenstein 2007) and their eating habits (Grunberg and Straub 1992).

Measuring Emotions

Emotions are a factor of external forces and are quite transient and short lived (Wilson 2001). Research, however, also shows that short-term emotions can have long-term impact, for example, there is enduring impact of transient emotions on economic decision making (Andrade and Ariely 2009). On the one hand, experiencing positive emotions serves to build enduring personal resources, ranging from physical and intellectual resources to social and psychological resources. On the other hand, experiencing negative emotions can lead to effects such as decline in learning and achievement (Pekrun 1992).

There are several methods to measure emotional impact. Galvanic skin sensitivity measurements or facial expression measurement methods have been used by many researchers. These methods, however, might induce artificiality in the live review setup. We choose non-intrusive methods that involve the use of self-reported scales. There are several scales that have been tested in the context of measuring emotions, such as the Self-Assessment Manikin (SAM), the Geneva Emotions Wheel (GEW) and the Positive Affect Negative Affect Scale (PANAS). We use the PANAS scale (Watson, Clark, and Tellegen 1988) which measures positive and negative emotional impact by asking respondents to rate 10 positive and 10 negative emotions each on a 5-point Likert scale.

The PANAS scale has been widely used and tested in a variety of emotion measurement scenarios similar to our research context: Rosenthal von der Pütten et al. (2013) investigated human's emotional reactions towards a robot, Mark et al. (2016) measured if sleep debt had an impact on the moods of students, Zhuang, Xie, and Lin (2017) measured effect of bright light therapy on emotions, Vuoskoski and Eerola (2015) examined if contextual information about a piece of music influences the emotions of listeners, Mekler and Hornbæk (2016) studied if eudaimonic user experiences led to positive emotions.

Grayscaling and Blurring Transformations

Prior research shows that color has the ability to influence a variety of human behaviors including the ability to categorize stimuli as positive or negative. Gohar (2008), for example, demonstrated that color is an important perceptual feature of six basic emotions: anger, disgust, fear, happiness, sadness, and surprise. Young et al. (2013), examined the ability of the color red to influence the identification of negative facial expressions, specifically anger. Sutton and Altarriba (2016), through their experiments revealed that red color was most commonly associated with negative emotion and emotion-laden words. Jeong, Biocca, and Kim (2011), reported that the presence of blood in games increased user's gore perception and aggressive thoughts. We chose blurring as another type of treatment, as there exists prior work that has proposed blurring of the original content, thereby allowing for the completion of the task at hand. Das et al. (2017), have used human computation to annotate blurred images. Dang, Riedl, and Lease 2018, have experimented with blurred images with varying degrees of blur in a simulated review environment. However, unlike these studies, we base our experiments on live review queues with human computation workers that do reviews as part of their work, thereby preserving external validity.

Method

Experiment Design and Setup

We use a pre-test post-test experimental design and test for two different types of interventions namely grayscaling and

blurring. The entire experiment lasted for four weeks. During the first two weeks of the experiment we took measurements from the regular setup (original untransformed) that served as our baseline (control). At the end of the two weeks, reviewers were given a questionnaire comprising two parts. The first part comprised the PANAS questionnaire and the second part comprised of questions about their review experience. During the following two weeks, reviewers reviewed all of the content in grayscale. For the grayscale treatment, reviewers had the option to switch to the original color mode for all or specific reviews. The blurring experiments were also setup in the same way as the grayscale experiments. The blur was set to 2.5% of the given images height. For example, an image with height 100px would have a blur effect of 2.5px. There is also a default value of 10px in the case were the images height cannot be determined. Reviewers could easily look at the image in its original form using a simple mouse-hover on the content. At the end of the treatment for two-weeks, reviewers answered the questionnaire comprising the PANAS scale. The treatment questionnaire had few additional questions about how likely the reviewers were to continue to use the treatment if given a choice and an open ended question to allow reviewers to share their experience. Table 1 gives a snapshot of the survey questionnaire for grayscale.

Research Ethics

Given the sensitive nature of the task, we implemented the following best practices to make sure our study was done in a respectful and most ethical manner possible. First, the study was completely anonymous, we refrained from collecting personally identifiable information and also provided instructions to reviewers to not include them in the open-ended text responses. Secondly, reviewers were made aware that any impact on quality during the course of the experiment would be attributed to the technical intervention introduced as part of the experiment and not on their individual performance. Thirdly, reviewers had the option to opt-out of the study at any point in time. Lastly, we did not collect or analyze demographic data such as gender, age group etc., as we did not plan to analyze demographic differences. We specifically omit this line of analysis to avoid potential incidental findings that may be used to discriminate such demography of reviewers in their review ability.

Participants

All study participants were content moderators who worked for the authors' organization. These moderators typically review content coming from live review queues. 76 reviewers participated in the grayscale study and 37 reviewers participated in the blurring study. A different set of reviewers participated in both studies to avoid recall and memory of the survey items.

Measurement

We used the following metrics to measure the impact of our experiments. First, we measured the emotional affect as measured by the PANAS scale. The PANAS scale measures

Table 1: Tests for significant differences between the experimental groups for AHT and accuracy (Grayscale)

This is a survey on the recent content you were reviewing. Please note that there are no right or wrong answers. We are not testing your memory, we are testing the review system. Your feedback will be much appreciated. Your name will NOT be recorded and the survey responses are anonymous. If you choose not to respond to the survey please click EXIT now, else please click on NEXT.

Indicate the extent to which you felt this way while reviewing the images on a scale of 1-5 — (1) Not at all; (2) Slightly; (3) Moderately (4) Very (5) Extremely

	(1)	(2)	(3)	(4)	(5)
Interested	○	○	○	○	○
Distressed	○	○	○	○	○
Excited	○	○	○	○	○
Strong	○	○	○	○	○
Upset	○	○	○	○	○
Guilty	○	○	○	○	○
Scared	○	○	○	○	○
Hostile	○	○	○	○	○
Enthusiastic	○	○	○	○	○
Proud	○	○	○	○	○
Irritable	○	○	○	○	○
Alert	○	○	○	○	○
Ashamed	○	○	○	○	○
Inspired	○	○	○	○	○
Nervous	○	○	○	○	○
Determined	○	○	○	○	○
Attentive	○	○	○	○	○
Jittery	○	○	○	○	○
Active	○	○	○	○	○
Afraid	○	○	○	○	○

If you are given an option to continue to review images in Grayscale, how likely are you to use the option?
Extremely likely — Highly likely — Neutral — Somewhat likely — Not at all likely

When reviewing images in Grayscale in the last few weeks, how often did you use the "view in color" feature?
Never — Rarely — Sometimes — Often — Always

Tell us about your experience reviewing the images in Grayscale.
<open ended>

change for 10 positive emotion and 10 negative emotions. To better understand the overall emotional impact of performing reviews we aggregate those scores into two meta scores: negative affect score is the sum of the 10 negative emotions, positive affect score is the sum of the 10 positive emotions. Secondly, alongside the scale based measurement, content reviewers provided responses to predefined open-ended interview questions that quizzed them on their experience with respect to the interventions. Lastly, we measured the mean review quality and average handling time (AHT) to ensure that such interventions do not have an adverse impact on business metrics such as quality and review time.

Results

Grayscale

Table 2 shows the results from a repeated measures ANOVA with a Greenhouse-Geisser correction. To ensure assumptions of repeated measures ANOVA are met, we tested the data for normality using KS test and sphericity using Mauchly's test and found that the data fulfilled these requirements. The test determined that AHT differed statistically significantly between color and grayscale conditions ($F(1, 4.115) = 17.804$, $p < 0.0005$) and the mean review accuracy showed no significant difference ($F(1, 0.003) = 0.071$, $p = .791$).

Table 2: Tests for significant differences between the experimental groups for AHT and accuracy (Grayscale).

	Exp. Group	Mean	F	Sig
Review Accuracy (n = 76)	Color	86.13%		
	Grayscale	87.79%	0.071	0.791
AHT (minutes) (n = 76)	Color	1.88		
	Grayscale	1.55	17.804	0.000

A Wilcoxon signed-rank test showed that the grayscale reviews led to a statistically significant increase in the positive affect of reviewers ($Z = -7.584$, $p < 0.001$). We use this non-parametric test as it does not assume normality in the data and given the affect scores were measured on an ordinal 5-pt likert scale. We further run pairwise tests between grayscale (treatment) and color (control) groups across the 20 constituent emotions for the PANAS scale. We observe statistically significant increases ($p < 0.05$) for the positive emotions: Attentive, Alert, Determined, Enthusiastic, Active and Proud for grayscale reviews. We did not observe significant differences in negative affect ($Z = -1.233$, $p < 0.217$). Table 3 shows the test for significance results. Figure 2 and 3 show the comparison across mean affect scores.

Feedback from Reviewers Reviewers that participated in the experiment had the opportunity to share their feedback about the experience. In addition to answering the PANAS survey we asked the reviewers to provide feedback on their experience on a post-experiment questionnaire. The

Table 3: Tests for significant differences between the experimental groups for Affect (Grayscale) ($p <= 0.05$)

	Exp. Group	Mean	Z	Sig
Positive Affect (n = 76)	Color	24.1		
	Grayscale	28.7	-7.584	0.000
Negative Affect (n = 76)	Color	17.1		
	Grayscale	19.0	-1.233	0.217

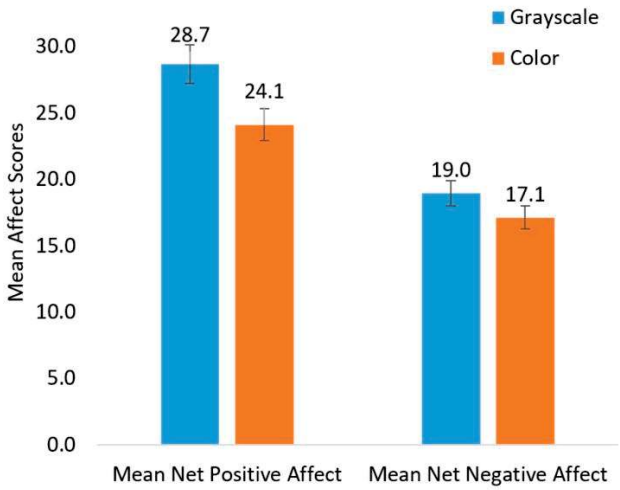

Figure 2: Comparison of Mean Affect scores (positive affect and negative affect) for grayscale and color.

feedback questionnaire asked the participants about how they felt about the intervention and how likely they were to adopt the intervention for their daily reviews if such a feature was made available. Several reviewers highlighted that grayscaling helped with better emotional balance and reduced the fright and trauma. Few reviewers also indicated that seeing less blood made them feel less nervous. The following are verbatim quotes from a sample of participants:

P[71] *"I felt more comfortable watching the videos in black and white . Somehow I find it even clearer, and easy on the eye , that I can focus less on cruelty . The level of emotional balance creates a more adequate environment to focus on precision while giving the verdict."*

P[27] *"In certain situations, it helps reviewer to minimize the fright and trauma while watching the cases."*

P[48] *"I felt better with the black and white as in black and white I could control my feelings more than colors content, felt less nervous and less disgusting."*

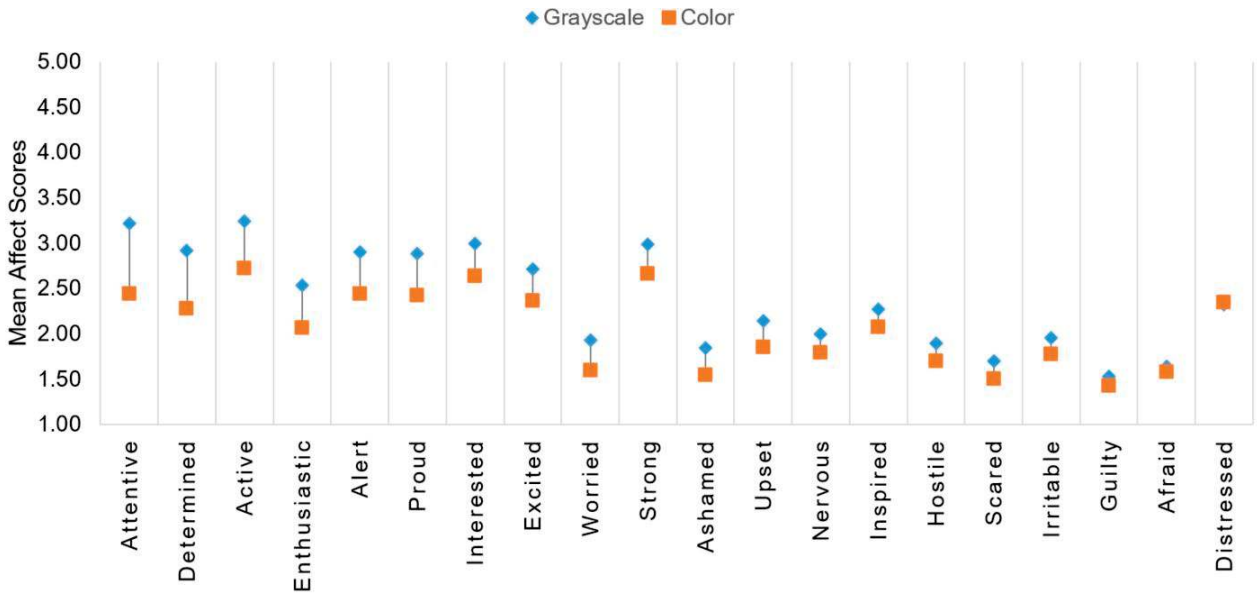

Figure 3: Mean affect scores across the constituent twenty emotions from the PANAS scale.

More than 70% of the reviewers indicated that they are likely to continue to use grayscale. Figure 4 shows the distribution of how likely the reviewers were to continue to use grayscale if given a choice. The remaining were concerned if changing to grayscale can make them less productive (we however, from our measurements note that there was no such drop and could in fact lead to productivity improvements). During the course of the grayscale experiment, reviewers had the option to revert to color for specific or all reviews. We followed up with the reviewers to measure how often reviewers engaged in reverting to color to make a review decision. About 1 in 5 reviewers never used the reverting option. Figure 5 shows the distribution of reviewers by frequency of reverting to color.

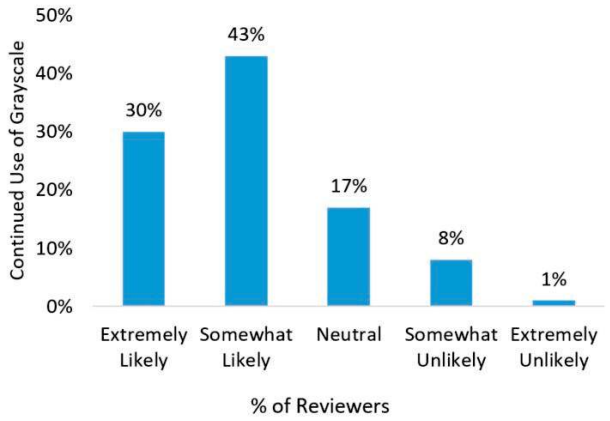

Figure 4: Distribution of reviewers based on likelihood of continued use of grayscale.

The remaining reviewers reverted to color less than 10 times during the entire course of review which typically comprised about 200 reviews. Reviewers did not mind an occasional switch to the color mode to make a review decision. For example, one of the reviewers provided the following explanation:

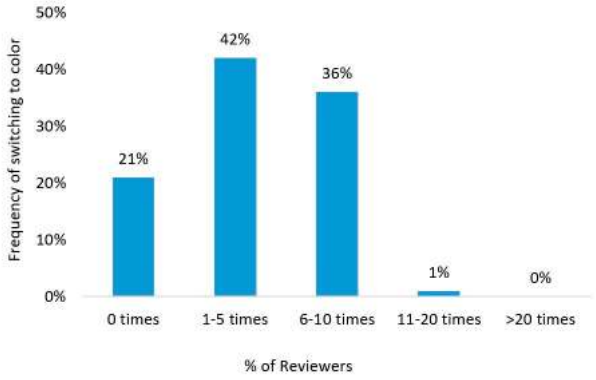

Figure 5: Distribution of reviewers by the frequency of selectively reviewing in color.

> **P[32]** *"I like the experience so much, even though I would have to change to color sometimes to fetch things more clearly, but in general it's a good experience to show mainly in grayscale and optionally change to color mode."*

Blurring

A repeated measures ANOVA (see Table 4) with a Greenhouse-Geisser correction determined that AHT did not

differ statistically significantly between regular (control) and blurring conditions (F(1, 0.238) = 1.984, p = .168) and the mean review accuracy showed no significant difference (F(1, 0.001) =0.192, p =.664). To ensure assumptions of repeated measures ANOVA are met, we tested the data for normality using KS test and sphericity using Mauchly's test and found that the data fulfilled these requirements.

Table 4: Tests for significance in means between the experimental groups for blurring intervention

	Exp. Group	Mean	F	Sig
Review Accuracy (n = 37)	Regular	91.50%		
	Blurring	91.19%	0.192	0.664
AHT (minutes) (n = 37)	Regular	2.14		
	Blurring	2.25	1.984	0.168

A Wilcoxon signed-rank test showed that the reviews in the blurred condition led to a statistically significant decrease in the positive affect of reviewers (Z = -3.275, p = 0.001). We further run these pairwise tests across the 20 constituent emotions for the PANAS scale. We observe statistically significant decreases ($p < 0.05$) for the positive emotions Interested (Z = -3.189, p = 0.001), Excited (Z = -3.343, p = 0.001), Strong (Z = -2.660, p = 0.008), Enthusiastic (Z = -2.926, p = 0.003), Proud (Z = -2.342, p = 0.019), Inspired (Z = -2.255 p = 0.024), Determined (Z = -3.160, p = 0.002) and Attentive (Z = -3.054, p = 0.002) for blur reviews. We did not observe significant differences in negative affect (Z = -.742, p = 0.458), however we did observe significant increase in one of the constituent emotions: 'Irritable' (Z = -2.099, p = 0.003). Table 5 shows the test for significance results. Figures 6 and 7 show the comparison across mean affect scores.

Table 5: Tests for significant differences between the experimental groups for Affect (Blurred) ($p <= 0.05$).

	Exp. Group	Mean	Z	Sig
Positive Affect (n = 37)	Regular	30.35		
	Blurring	22.84	-3.28	0.001
Negative Affect) (n = 37)	Regular	14.19		
	Blurring	15.68	-.742	0.458

Feedback from Reviewers Overall, reviewers shared negative sentiment about the blurring review experience. 78% of the reviewers indicated that they are unlikely to continue to use blurring. Figure 8 shows the distribution of how likely the reviewers were to continue to use blurring if given a choice. Similar to the grayscale task, reviewers that participated in the experiment had the opportunity to share

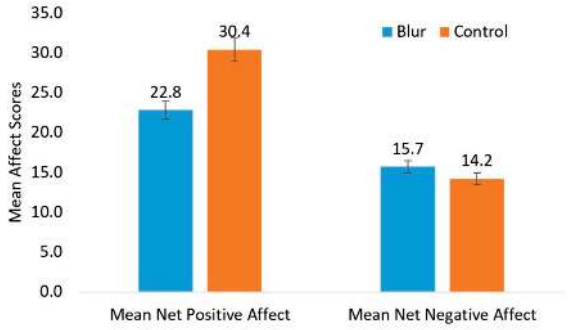

Figure 6: Comparison of Mean Affect scores (positive affect and negative affect) for blurred and control.

their feedback about the experience. Several reviewers also highlighted that the blurred images had a strain on their eyes.

P[8] *"I felt much more tired. The blurred images made my eyes strained."*

P[20] *"This was not a good experience. aside the fact that it took longer to do the reviews, also I felt that my focus was all over the place and also due to the blur my eyes could not focus and it gave me a headache."*

Few reviewers were also concerned about review quality and felt confused due to lack of clarity with the content.

P[15] *"Its make me feel dizzy and most of time impacted on the verdict I can't give the right verdict unless I remove the blurring."*

Reviewers also expressed usability issues with the use of blur. For example a reviewer highlighted that content was not clear at all and another reviewer indicated that the blurring effect slowed down the page loading on the review tool.

P[31] *"Very nervous to fail the review since some of the content was not clear at all. It was a bit confusing and stressful."*

P[11] *"It is very inconvenient and double the work. Not only did it slow down my workflow, it made the loading of the pages slower as well. It was also very time consuming to roll over each image just to see what it was."*

Discussion

Implications

Across the two interventions tested we see that the grayscale approach worked best in achieving the goals of our experiment of enhancing emotional affect without compromising business and operational metrics. It is however, imperative

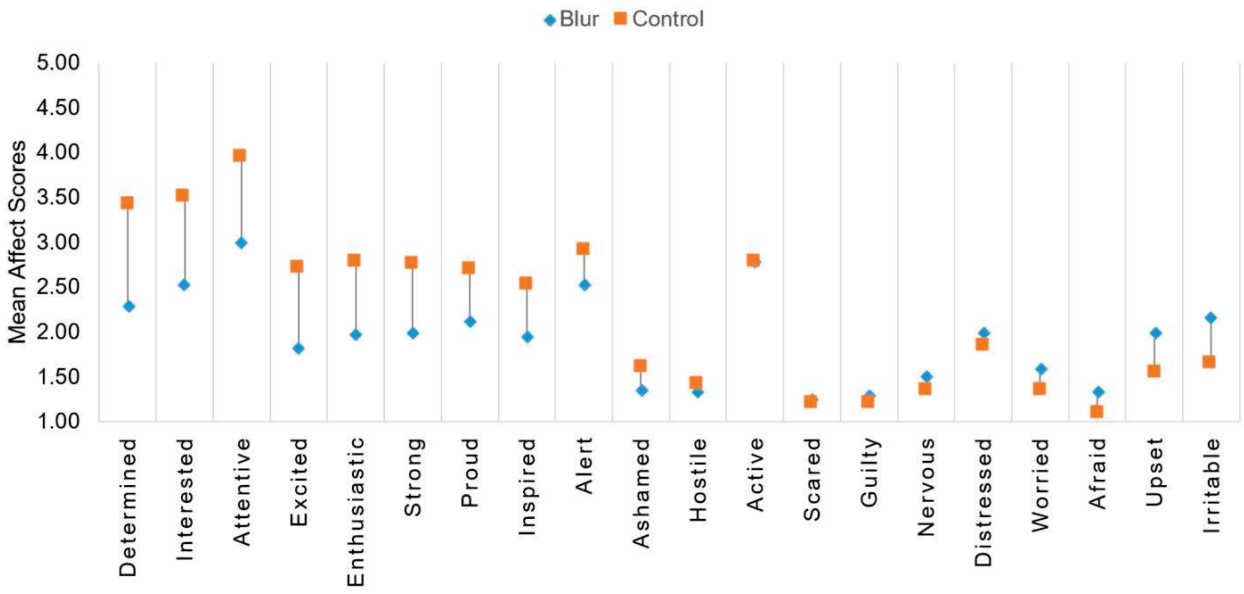

Figure 7: Comparison of affect scores across grayscale and color- break down by 20 constituent emotions.

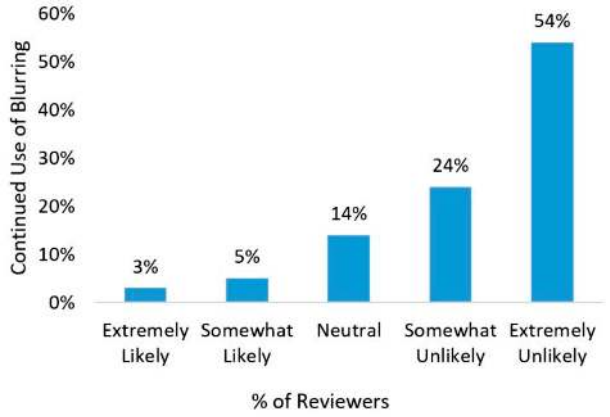

Figure 8: Distribution of reviewers based on likely continued use of blurring.

that the choice of using a specific intervention should be left with the reviewers. In our setup, post successful results from the experiment, we have provided the grayscale feature as an option for the reviewers to choose from and we refrained from making this a mandatory change. We observed that 70% of the reviewers that participated in the study have started using grayscale on a regular basis.

As identified from the blurring study, an intervention that looks promising must also be well-designed to seamlessly fit into a reviewer's process. Adequate considerations need to be made to ensure the new intervention does not cause usability issues such as more number of clicks or increase in loading times. There is a need to engage user experience designers and engineers to implement such interventions into the review tools.

Our important contribution to this area of research is the testing with actual reviewers who work on content moderation reviews as part of their daily jobs. Below is a quote from a reviewer from the study who captures the essence of the mindset of many of these reviewers who take pride in making the internet a better place. As researchers and tool developers work on identifying new mechanisms to help with the well-being of the reviewers, it is important to involve the reviewers to provide feedback on the solutions being developed.

> **P[52]***"I feel like we are the E-Special Force that look after the whole world and protect them from watching and viewing disturbing content and at the same time we are fighting against those who share and upload those content that terrifies people. This job is great responsibility and it makes me feel proud."*

Limitations

We provide an approach that helps introduce and measure stylistic interventions on live review queues. The applicability of our findings is limited to the content types seen in the review queues we studied. While emotions are transient and short-lived, there is research that shows that emotions can also have a long-term impact. For practical reasons we use a two week study window for our treatment. The long-term efficacy and impact would also have to be investigated. Given our study was fully anonymous and did we not collect any demographic information, we are unable to compare differences in emotional affect across demographics. This is a known limitation and we chose this route to avoid unintended incidental findings that could potentially be used during hiring decisions of human computation workers.

Future Research Directions

Testing the impact of several other interventions such as masking specific colors (for example, changing all red to green), selective blurring, artistic transformations, different shades of grayscaling, etc. are all future research opportunities. Extending research in the domain of privacy controls (for example, (Hasan et al. 2018)) that have explored the concept of obfuscating sensitive aspects of images could be invaluable for human computation tasks.

Our research also demonstrates the value in measuring affect which is an indicator of well-being. Identifying simple, reliable long-term longitudinal measurement scales that can be deployed on a continuous basis can be another area for future research. Our research provides an approach that researchers in the area of human computation could apply to learn the effects of various stylistic, artistic or other types of interventions that mitigate emotional impact. While there are several scales to measure emotional impact we demonstrate the use of the PANAS scale to be able to track emotional impact within the constraints of testing in a live review setup.

Conclusion

We find that simple stylistic transformations can provide an easy to implement solution to significantly reduce the emotional impact of manual content reviews. In this paper we systematically quantify the emotional impact on human reviewers when reviewing images using different stylistic interventions and test the effectiveness of such transformations in improving the emotions experienced by reviewers while preserving review accuracy and time required for reviews. One of our key findings suggests reviewing content in grayscale improved positive affect of reviewers while reviewing the most violent and extreme images in a statistically significant manner. Blurring the content, however, entirely had a further negative impact on emotional affect of reviewers, and in particular increasing the irritability of reviewers. Overall our study provides evidence that the negative emotional impact on reviewers can be mitigated through the use of image transformations such as grayscale. By designing interventions that help content moderators better cope with their work, we seek to minimize possible risks associated with moderating difficult content, whilst preserving the accuracy and handling time of such reviews.

References

Andrade, E. B., and Ariely, D. 2009. The enduring impact of transient emotions on decision making. *Organizational Behavior and Human Decision Processes* 109(1):1–8.

Canegallo, K. 2019. Meet the teams keeping our corner of the internet safer. The Keyword.

Chen, A. 2014. The laborers who keep dick pics and beheadings out of your facebook feed. *Wired* 23:14.

Dang, B.; Riedl, M. J.; and Lease, M. 2018. But who protects the moderators? the case of crowdsourced image moderation. *arXiv preprint arXiv:1804.10999*.

Das, A.; Agrawal, H.; Zitnick, L.; Parikh, D.; and Batra, D. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* 163:90–100.

Deniz, O.; Serrano, I.; Bueno, G.; and Kim, T.-K. 2014. Fast violence detection in video. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, 478–485. IEEE.

Gao, Y.; Liu, H.; Sun, X.; Wang, C.; and Liu, Y. 2016. Violence detection using oriented violent flows. *Image and vision computing* 48:37–41.

Ghoshal, A. 2017. Microsoft sued by employees who developed ptsd after reviewing disturbing content. The next web.

Gillespie, T. 2017. Governance of and by platforms. SAGE Handbook of Social Media (Jean Burgess, Thomas Poell, and Alice Marwick, eds).

Gohar, N. K. 2008. *Diagnostic colours of emotions.* Ph.D. Dissertation, University of Sydney.

Grunberg, N. E., and Straub, R. O. 1992. The role of gender and taste class in the effects of stress on eating. *Health Psychology* 11(2):97.

Hasan, R.; Hassan, E.; Li, Y.; Caine, K.; Crandall, D. J.; Hoyle, R.; and Kapadia, A. 2018. Viewer experience of obscuring scene elements in photos to enhance privacy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 47. ACM.

Jeong, E.-J.; Biocca, F. A.; and Kim, M.-K. 2011. Realism cues and memory in computer games: Effects of violence cues on arousal, engagement, and memory. *Journal of Korea Game Society* 11(4):127–142.

Krause, T., and Grassegger, H. 2016. Inside facebook. *Süddeutsche Zeitung* 15.

Mark, G.; Niiya, M.; Reich, S.; et al. 2016. Sleep debt in student life: Online attention focus, facebook, and mood. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 5517–5528. ACM.

Mekler, E. D., and Hornbæk, K. 2016. Momentary pleasure or lasting meaning?: Distinguishing eudaimonic and hedonic user experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4509–4520. ACM.

Pekrun, R. 1992. The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators. *Applied Psychology* 41(4):359–376.

Roberts, S. T. 2016. Commercial content moderation: Digital laborers' dirty work.

Roberts, S. T. 2017. Content moderation. In Schintler, L. A., and McNeely, C. L., eds., *Encyclopedia of Big Data*. Springer International Publishing. 1–4.

Roberts, S. T. 2018. Digital detritus:'error'and the logic of opacity in social media content moderation. *First Monday* 23(3).

Rojas-Galeano, S. 2017. On obstructing obscenity obfuscation. *ACM Transactions on the Web (TWEB)* 11(2):12.

Rosenthal von der Pütten, A. M.; Krämer, N. C.; Hoffmann, L.; Sobieraj, S.; and Eimler, S. C. 2013. An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics* 5(1):17–34.

Schmidt, A., and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10.

Sutton, T. M., and Altarriba, J. 2016. Color associations to emotion and emotion-laden words: A collection of norms for stimulus construction and selection. *Behavior research methods* 48(2):686–728.

Vohs, K. D.; Baumeister, R. F.; and Loewenstein, G. 2007. *Do Emotions Help or Hurt Decisionmaking?: A Hedgefoxian Perspective*. Russell Sage Foundation.

Vuoskoski, J. K., and Eerola, T. 2015. Extramusical information contributes to emotions induced by music. *Psychology of Music* 43(2):262–274.

Watson, D.; Clark, L. A.; and Tellegen, A. 1988. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology* 54(6):1063.

Young, S. G.; Elliot, A. J.; Feltman, R.; and Ambady, N. 2013. Red enhances the processing of facial expressions of anger. *Emotion* 13(3):380.

Zhuang, Y.; Xie, K.; and Lin, Y. 2017. Effect of bright light therapy for depression. In *2017 14th China International Forum on Solid State Lighting: International Forum on Wide Bandgap Semiconductors China (SSLChina: IFWS)*, 109–112. IEEE.