Testing Grayscale Interventions to Reduce Negative Emotional Impact on Manual Reviewers

Sowmya Karunakaran
sowmyakaru@google.com
Google Ireland

Rashmi Ramakrishnan
raramakrishnan@google.com
Google Singapore

1. 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 [1]. While machines and technology play a critical role in content moderation, there continues to be a need for manual reviews where human judgement is required in interpreting borderline cases as well as generation of ground truth for ML models. It is known, however, that such manual reviews could be emotionally challenging.

2. Background

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. 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.

3. Method

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 [2] as it was successfully used to measure emotional response in similar research contexts. The PANAS scale measures positive and negative emotional impact by asking respondents to rate 10 positive and 10 negative emotions each on a 5-point Likert scale.

We use a pretest-posttest experimental design. We refrain from collecting any personally identifiable information to keep the study fully anonymous. Reviewers had the option to opt-out of taking the PANAS survey. Alongside the scale based measurement, content reviewers provided responses to predefined interview questions that quizzed them on their experience with respect to the interventions. We also measure review quality and time duration of reviews to ensure that such interventions do not have an adverse impact on business metrics.

4. Conclusion

We find that simple stylistic transformations can provide an easy to implement solution to significantly reduce the emotional impact of manual content reviews.

References

1. Canegallo, K. (2019). Meet the teams keeping our corner of the internet safer. [Blog] *The Keyword*.
2. Watson, D., Clark, L. A., & 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.