Towards Effective Crowd-Powered Online Content Moderation

Danula Hettiachchi
The University of Melbourne
Melbourne, Australia
danula.hettiachchi@unimelb.edu.au

Jorge Goncalves
The University of Melbourne
Melbourne, Australia
jorge.goncalves@unimelb.edu.au

ABSTRACT

Content moderation is an important element of social computing systems that facilitates positive social interaction in online platforms. Current solutions for moderation including human moderation via commercial teams are not effective and have failed to meet the demands of growing volumes of online user generated content. Through a study where we ask crowd workers to moderate tweets, we demonstrate that crowdsourcing is a promising solution for content moderation. We also report a strong relationship between the sentiment of a tweet and its appropriateness to appear in public media. Our analysis on worker responses further reveals several key factors that affect the judgement of crowd moderators when deciding on the suitability of text content. Our findings contribute towards the development of future robust moderation systems that utilise crowdsourcing.

CCS CONCEPTS

• **Information systems → Crowdsourcing**; *Collaborative and social computing systems and tools.*

KEYWORDS

crowdsourcing, content moderation, social computing, twitter sentiment analysis

ACM Reference Format:
Danula Hettiachchi and Jorge Goncalves. 2019. Towards Effective Crowd-Powered Online Content Moderation. In *31ST AUSTRALIAN CONFERENCE ON HUMAN-COMPUTER-INTERACTION (OZCHI'19), December 2–5, 2019, Fremantle, WN, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3369457.3369491

1 INTRODUCTION

With the volume of online user generated content growing steadily over the past years, moderating this content has become an important challenge. Popular social media networks acknowledge that moderation is essential to ensure the safety online and have set out clear community guideline or policies to determine what content will be potentially removed from platforms (*e.g.,* Facebook Community Standards[1], Twitter Media Policy[2], and YouTube Community Guidelines[3]). Such platforms use dedicated human content moderators [25] and advanced machine learning techniques for content moderation. However, even with relatively large moderation teams, these platforms often fail to cater for the growing demand[4]. Apart from popular social media networks, most current commercial moderation methods use simple approaches such as black-lists and regular expressions. These methods tend to fail and scale poorly when assessing more elusive content such as hate speech or derogatory (language which attacks an individual or a group but not hate-speech) in large volumes populated by users from diverse backgrounds. Thus, moderating online user generated content requires a highly versatile and scalable solution.

Crowdsourcing provides an economical and effective way to reach a large number of online workers [3, 18, 19]. As the collective judgement of crowd workers is often comparable or superior to the automated approaches [1, 11, 14], crowdsourcing presents itself as an effective approach for content moderation [23]. In this work, we use the wisdom of the crowd to create a more robust content moderation mechanism. We conducted a study on Amazon Mechanical Turk (MTurk)[5] where we asked participants to rate if a given tweet is appropriate for a general audience. 28 workers provided 2,400 labels along with justifications for their judgements. We show that human annotation with regard to the appropriateness of content is also correlated with the sentiment of the tweet evaluated by a state-of-the-art automated sentiment classifier. We further explore different motivations a moderator would have to mark content as inappropriate.

2 RELATED WORK

2.1 Moderating User Content

Moderating online user generated content has always been recognised as a highly challenging task [22]. Nobata et al. [24] summarises several concerns that can increase the complexity in the moderation process. First, users who create inappropriate content aim to avoid moderation by intentionally obfuscating words or phrases (*e.g.,* through replacing letters with numbers or special characters). Second, racial or minority insults can vary depending on many factors like the context, the targeted group, and the locality. Specific phrases and words also evolve over time. Thus, using a static mechanism is not sufficient. Third, some inappropriate content may still be free of grammatical errors and sound very eloquent. This makes it difficult, even for an experienced moderator,

OZCHI'19, December 2–5, 2019, Fremantle, WN, Australia

ACM ISBN 978-1-4503-7696-9/19/12…$15.00
https://doi.org/10.1145/3369457.3369491

[1] https://www.facebook.com/communitystandards/objectionable_content
[2] https://help.twitter.com/en/rules-and-policies/media-policy
[3] https://www.youtube.com/yt/about/policies/#community-guidelines
[4] https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways
[5] https://www.mturk.com

to spot them at first sight. Fourth, the content may span into several sentences. Automated methods often fail to make a judgement by inferring meaning from a sequence of sentences. Finally, linguistic features like sarcasm, metaphors and irony could result in user content being inaccurately flagged as inappropriate content.

Research has shown it is effective to incorporate different natural language processing techniques such as sentiment analysis, text normalization with machine learning methods to achieve better automated moderation [24, 26]. However, such implementations are mostly limited to specific types of inappropriate content such as profanity. Moderating non-text-based content such as images and videos could be more challenging than text-based content. Specifically, computer vision research has used deep learning techniques to identify the sentiment of an image [29]. Regardless, model-based moderation requires a high-quality annotated dataset which needs to be updated constantly to keep up with the constant changes in online communities.

Sentiment Analysis, particularly on Twitter data, has been extensively researched [20]. Apart from the main text included in the tweet, many other parameters like hashtags, emoticons, number of retweets are known to be useful in sentiment analysis. Several studies have used crowdsourcing for sentiment analysis specifically to create a base sentiment model and to further enhance the recent and misclassified data in a real-time stream [28].

2.2 Crowdsourcing for Content Moderation

Crowdsourcing has been successfully utilised to harness large volumes of human judgements for tasks that are traditionally complex for machines [3, 10]. For example, crowdsourcing can be used to detect fraudulent product reviews that appear authentic and written with the intent to mislead [14].

Ghosh et al. [5] first proposed a framework for using crowdsourcing for moderating user generated content. They presented an efficient algorithm that can detect abusive content using the identity of a single trustworthy contributor. However, their validation is limited to a simulation and does not include a deployment in an actual crowdsourcing platform. In a study that examines abusive behaviour on Twitter, Founta et al. [4] used crowdsourcing to create a collection of tweets with abuse-related labels. Such data sets can be used as training data for machine learning models that detect inappropriate content. Similarly, Chatzakou et al. [2] used crowdsourcing to tag Twitter user profiles as either 'bully', 'aggressor', or 'spammer'. They used crowdsourced data as ground-truth to train and evaluate their proposed method.

Apart from online crowdsourcing, crowd moderation has also been used in situated crowdsourcing [8, 15, 16]. In practice, one major challenge in using crowdsourcing for content moderation is the demand for real-time processing. In public displays, moderation delay is known to significantly reduce the number of user generated content [13].

In our work, we deploy a study on MTurk that involves crowd workers in the moderation process and demonstrates the feasibility of using crowdsourcing for online content moderation. Using reasons provided by workers along with their labels, we further explore factors that people consider to arrive at their judgement.

3 STUDY

Our dataset included a total of 8,665 tweets extracted from Twitter with the keyword 'Obama' with a period of 20 days. The keyword selected was informed by the need to ensure we obtain tweets that come from a diverse set of users, tweets that potentially contain various types of inappropriate content like profanity, and tweets that are possibly linked to personal beliefs or preferences. The dataset did not include any duplicate tweets and we also removed any links or images in tweets.

First, we used Vader [17], a well established lexicon-based sentiment analysis tool to gauge the sentiment of each tweet. Vader provides three sentiment scores: negative, positive and compound. Using the compound sentiment score, we curated a dataset of 300 tweets to be used in the crowdsourcing study. Compound sentiment score is a decimal value that ranges from negative (-1) to positive (+1). We divided the tweets into 20 buckets with each accounting for a score range of 0.1 (*e.g.*,0 to 0.1). Then from each bucket we randomly selected 15 tweets for the crowdsourcing dataset.

The crowdsourcing study was deployed on Amazon Mechanical Turk. In each HIT (Human Intelligent Task), we presented workers with a tweet and asked them if the given tweet is suitable for a general audience (*i.e.*,to appear on a TV show). Workers categorised the tweet as either 'Appropriate', 'Inappropriate' or 'Unsure' and then optionally provided a reason for the selection. For each tweet, we obtained answers from 8 different workers and the aggregated rating for each tweet was calculated using the following formula.

A - Number of Appropriate labels
I - Number of Inappropriate labels
U - Number of Unsure labels

$$\text{Aggregated Rating} = A \times (+1) + U \times (0) + I \times (-1)$$

Based on our evaluation, the completion of each individual task (moderating a tweet and providing any comments) takes around 1.5 minutes. Based on the highest state minimum wage of the US $11.50, we paid $0.30 for each tweet. The amount we paid for a worker is comfortably above the average pay one would receive by completing regular tasks in MTurk. The study is approved by the Ethics committee of our university.

4 RESULTS

A total of 28 crowd participants completed the study. All participants are from the US and worker age range from 25 to 62 years ($M = 36.1$, $SD = 8.7$). For 300 tweets, we obtained 2,400 individual labels in total. Figure 1 shows the distribution of tweets after calculating the aggregated rating for each tweet. We see a large number of tweets that has been approved by all the assigned participants resulting in an aggregated rating of 8.

There are 8 participants who provided more than 200 labels. For each of these participants considering tweets that had at least 1 unsure or inappropriate label, we calculated the deviation of their ratings from the aggregated rating. Figure 2 shows the mean deviation scores across the 8 participants and for certain workers, we note a considerable difference in their lenience towards accepting or rejecting content.

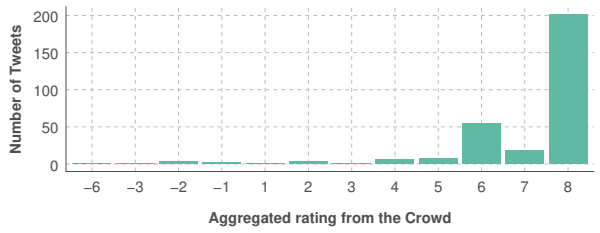

Figure 1: Variation in the number of tweets for each aggregated rating score

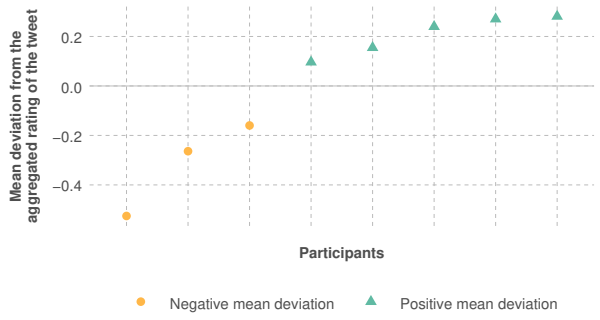

Figure 2: Difference in mean deviation from aggregated rating for participants who labelled more than 200 tweets

4.1 Sentiment and Moderation

As shown in Figure 3, tweets with a positive compound sentiment score are more likely to be marked as appropriate by the crowd. Here, the final aggregated label is classified as appropriate if the aggregated rating is greater than 6 and classified as inappropriate otherwise. A Wilcoxon rank sum test further confirms that there is a significant difference in compound sentiment score among the two groups of tweets based on the crowd labelling ($W = 10295$, $p = 0.02$).

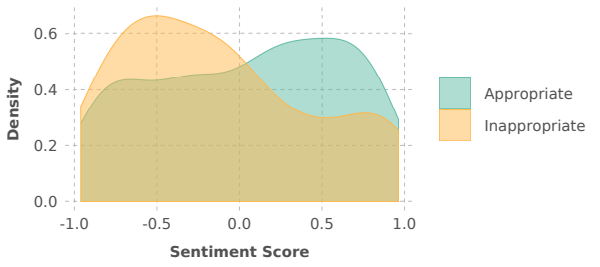

Figure 3: Density plot for crowd label for tweets

4.2 Responses from Crowd Workers

We examined the reasons provided by participants along with the label as the response for each HIT. Workers rarely provided a reason after labelling a tweet as appropriate. There were 125 responses where workers provided a reason along with a inappropriate label and 34 responses with a unsure label whereas only 3 responses contained a reason for an appropriate label. We examined the reasons provided by workers along with inappropriate and unsure labels to gain a better understanding of the moderation process and the factors influenced the participants judgement.

The authors carefully examined and categorised each reason under four categories: profanity and hate-speech, language issues, opinion related, off-topic, and other.

4.2.1 Profanity and Hate-Speech. Workers seem to be quite confident about the labels provided under profanity and hate-speech. These tweets mainly contained swear words or words that are generally considered rude or offensive, defamatory claims directed at a particular person, and terms or phrases that express prejudice against a particular group. We also noticed that corresponding tweets had a lower aggregated rating, meaning that they were labelled as inappropriate by a majority of workers.

> P04: *"Message is aggressive, uses profanity."*

> P18: *"Inappropriate language"*

4.2.2 Language Issues. Tweets that lack the overall quality in terms of coherence, grammar and spelling were also penalised by many workers. Workers often used terms such as 'nonsensical', 'incoherent', and 'unreadable' to describe these tweets.

> P11: *"the tweet is so misspelled it shouldn't be shown on TV"*

4.2.3 Opinion Related. Certain workers labelled tweets as inappropriate due to their personal opinion. These tweets were often marked as appropriate by a majority of workers. For example, the following tweets were labelled as inappropriate by a single crowd worker where as 7 workers labelled them as appropriate. The reasons provided by the workers to justify the inappropriate label suggest an influence of personal opinion.

> Tweet: "There are people out there who have been stockpiling weapons for decades and they are very eager to get to use them in their Hollywood daydream scenario where Obama comes for their guns in the night. Mass shootings will seem quaint."
>
> P11: *"Seems insulting to the conservative gun owners you can't put this on TV for a general audience"*

> Tweet: "Under President Obama, college graduates were forced to move back home due to the job market. Today, youth unemployment is at a 52 year low!"
>
> P18: *"Strongly anti-Obama"*

4.2.4 Off-topic. A number of workers labelled tweets as inappropriate as they were not relevant to the topic 'Obama' or contained advertisements.

> P07: *"Advertisement, nothing to do with the topic."*

4.2.5 Other. Reasons that did not fall into above categories represent tweets that mention sensitive topics and that contains inaccurate facts or details.

> P18: *"Unsubstantiated defamatory claim"*

4.2.6 Reasons for Unsure Labels. We further examine the reasons provided along with unsure labels. Some workers were uncertain if a particular topic is sensitive or not for the given moderation condition.

> P04: *"Not sure because it mentions 'weed'. Although, it is a direct quote, so it's possible that it would be used."*

In certain cases, workers were also finding it difficult to determine if a particular tweet qualifies to be marked as inappropriate. This is mainly because, either they could not understand the content properly to make a judgement or they were uncertain if a particular term or expression should be deemed appropriate or not.

> P21: *"I am not sure the tone/language of this tweet is appropriate for a general audience."*

> P13: *"I'm not sure mainly because it is grammatical nonsense and rambling"*

> P23: *"It doesn't make sense."*

5 DISCUSSION

Crowdsourcing can be utilised to moderate online user generated content in multiple ways. First, as we show in our study and as suggested by Ghosh et al. [5], crowd moderation could be implemented as a standalone moderation mechanism. Second, as demonstrated in prior work [2, 4], crowd workers can contribute to create accurate and rich training labels for supervised learning based automated approaches. Third, by combining crowdsourcing and automated approaches, a human-in-the-loop moderation system could be implemented. In such a system, the content will be reviewed by crowd workers, when automated approaches fail to make a confident judgement.

Our analysis of participant comments reveals interesting insights that can help create an effective crowdsourcing mechanism for content moderation. We note participants often labelled unsure when they found it difficult to decide on borderline content. This could be improved by providing specific guidelines or illustrative examples for participants. Many crowdsourcing studies have highlighted the importance in providing clear instructions to obtain high quality data [7, 9, 21, 27]. Further, the use of comparative judgements in place of discrete labels can lead to more reliable answers [12].

From the justifications we received from the crowd workers engaged in the study, we also notice that individual opinion could have a considerable impact on the moderation process. We also see in Figure 2 that certain workers lean more towards labelling content either as appropriate or inappropriate compared to the rest. As we obtain multiple labels for each item, the influence of personal opinion is expected to be reduced when using crowdsourcing. However, this is also valid for dedicated moderation teams and can have broader implications when only one person reviews content. Therefore, it is important to set out clear policies, educate and train commercial content moderators to limit the influence of personal opinion. In addition, it is also important to take steps towards ensuring that a diverse set of people contribute to the moderation process.

Any moderation process that involves humans has many ethical considerations such as exposure to extreme and explicit content [6]. Human moderators who work in commercial moderation teams have often suffered from negative effects of prolonged exposure to such content. This is also one of the main reasons behind low retention rates in employment related to moderation [25]. We can argue that crowdsourcing is relatively better than commercial moderation as crowd workers have the freedom to leave a task at any point. However, it is essential to carefully consider and take appropriate measures to limit the potential negative impacts of content moderation that involves humans.

5.1 Limitations

We acknowledge several limitations of our study. First, in our online deployment, we only received labels from 28 workers. This is partly due to the nature of typical crowd market places where it is not straightforward to manage the task assignment. This limits our ability to analyse any impact of participant attributes such as age, education level and political leaning on the ratings. Second, although we created a well-balanced dataset using sentiment score, the result set has a large proportion of appropriate tweets that has little value for our analysis. Third, our study does not dynamically moderate the tweets which is required to enable real world implementation.

6 CONCLUSION AND FUTURE WORK

Using a dataset that contains tweets on a potentially divisive topic, we show that it is feasible to use crowdsourcing for content moderation. From the results of our study, we also establish a strong relationship between the sentiment of a tweet and the appropriateness of the tweet to appear in public media. However, numerous challenges emerge when utilising independent crowd workers for moderation through online platforms. We uncover and discuss several factors that influence the judgement of an individual related to classifying text content as appropriate or not for a general audience.

A future study on the impact of worker demographics and other attributes on the moderation process could pave the way to reduce the bias that could potentially obstruct fair moderation. Moderation delay also has a significant impact on the moderation process. Therefore, a future study that explores dynamic crowd moderation could provide further insights on how to utilise crowdsourcing for moderation.

ACKNOWLEDGMENTS

We thank Kening Wu for her contributions to the study, particularly during the data collection process.

REFERENCES

[1] Daren C. Brabham. 2008. Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence* 14, 1 (2008), 75–90. https://doi.org/10.1177/1354856507084420

[2] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. In *Proceedings of the 2017 ACM on Web*

Science Conference (WebSci '17). ACM, New York, NY, USA, 13–22. https://doi.org/10.1145/3091478.3091487

[3] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 238–247. https://doi.org/10.1145/2736277.2741685

[4] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Twelfth International AAAI Conference on Web and Social Media*. AAAI Press.

[5] Arpita Ghosh, Satyen Kale, and Preston McAfee. 2011. Who Moderates the Moderators?: Crowdsourcing Abuse Detection in User-generated Content. In *Proceedings of the 12th ACM Conference on Electronic Commerce (EC '11)*. ACM, New York, NY, USA, 167–176. https://doi.org/10.1145/1993574.1993599

[6] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

[7] Jorge Goncalves, Denzil Ferreira, Simo Hosio, Yong Liu, Jakob Rogstadius, Hannu Kukka, and Vassilis Kostakos. 2013. Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 753–762. https://doi.org/10.1145/2493432.2493481

[8] Jorge Goncalves, Simo Hosio, Denzil Ferreira, and Vassilis Kostakos. 2014. Game of Words: Tagging Places Through Crowdsourcing on Public Displays. In *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS '14)*. 705–714. https://doi.org/10.1145/2598510.2598514

[9] Jorge Goncalves, Simo Hosio, Jakob Rogstadius, Evangelos Karapanos, and Vassilis Kostakos. 2015. Motivating participation and improving quality of contribution in ubiquitous crowdsourcing. *Computer Networks* 90 (2015), 34 – 48. https://doi.org/10.1016/j.comnet.2015.07.002

[10] Jorge Goncalves, Simo Hosio, Niels van Berkel, Furqan Ahmed, and Vassilis Kostakos. 2017. CrowdPickUp: Crowdsourcing Task Pickup in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 51 (Sept. 2017), 22 pages. https://doi.org/10.1145/3130916

[11] Jorge Goncalves, Hannu Kukka, Iván Sánchez, and Vassilis Kostakos. 2016. Crowdsourcing Queue Estimations in Situ. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1040–1051. https://doi.org/10.1145/2818048.2819997

[12] Sian Gooding, Ekaterina Kochmar, Advait Sarkar, and Alan Blackwell. 2019. Comparative judgments are more consistent than binary classification for labelling word complexity. In *Proceedings of the 13th Linguistic Annotation Workshop*. Association for Computational Linguistics, Florence, Italy, 208–214. https://doi.org/10.18653/v1/W19-4024

[13] Miriam Greis, Florian Alt, Niels Henze, and Nemanja Memarovic. 2014. I Can Wait a Minute: Uncovering the Optimal Delay Time for Pre-moderated User-generated Content on Public Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 1435–1438. https://doi.org/10.1145/2556288.2557186

[14] Christopher Glenn Harris. 2012. Detecting deceptive opinion spam using human computation. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press.

[15] Simo Hosio, Jorge Goncalves, Vassilis Kostakos, and Jukka Riekki. 2015. Crowdsourcing Public Opinion Using Urban Pervasive Technologies: Lessons From Real-Life Experiments in Oulu. *Policy & Internet* 7, 2 (2015), 203–222. https://doi.org/10.1002/poi3.90

[16] Simo Hosio, Jorge Goncalves, Vili Lehdonvirta, Denzil Ferreira, and Vassilis Kostakos. 2014. Situated Crowdsourcing Using a Market Model. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, NY, USA, 55–64. https://doi.org/10.1145/2642918.2647362

[17] Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Eighth International AAAI Conference on Web and Social Media*. AAAI Press.

[18] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[19] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[20] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG!. In *Fifth International AAAI Conference on Web and Social Media*. AAAI Press.

[21] Anand Kulkarni, Philipp Gutheim, Prayag Narula, David Rolnitzky, Tapan Parikh, and Björn Hartmann. 2012. MobileWorks: Designing for Quality in a Managed Crowdsourcing Architecture. *IEEE Internet Computing* 16, 5 (Sep. 2012), 28–35. https://doi.org/10.1109/MIC.2012.72

[22] Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 543–550. https://doi.org/10.1145/985692.985761

[23] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31, 2 (2014), 317 – 326. https://doi.org/10.1016/j.giq.2013.11.005

[24] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 145–153. https://doi.org/10.1145/2872427.2883062

[25] Sarah T. Roberts. 2016. *Commercial Content Moderation: Digital Laborers' Dirty Work*. Peter Lang, Bern, Switzerland.

[26] Anna Schmidt and Michael Wiegand. 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.

[27] Oksana Tokarchuk, Roberta Cuel, and Marco Zamarian. 2012. Analyzing Crowd Labor and Designing Incentives for Humans in the Loop. *IEEE Internet Computing* 16, 5 (Sep. 2012), 45–51. https://doi.org/10.1109/MIC.2012.66

[28] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations*. Association for Computational Linguistics, 115–120.

[29] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. In *Twenty-ninth AAAI Conference on Artificial Intelligence*. AAAI Press.