Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions

Lora Aroyo
Google
New York, NY
l.m.aroyo@gmail.com

Lucas Dixon
Jigsaw
Paris, France
ldixon@google.com

Olivia Redfield
Jigsaw
New York, NY
orhinehart@google.com

Rachel Rosen
Jigsaw
New York, NY
rachelrosen@google.com

Nithum Thain
Jigsaw
Montreal, Canada
nthain@google.com

ABSTRACT

Discussing things you care about can be difficult, especially via online platforms, where sharing your opinion leaves you open to the real and immediate threats of abuse and harassment. Due to these threats, people stop expressing themselves and give up on seeking different opinions. Recent research efforts focus on examining the strengths and weaknesses (e.g. potential unintended biases) of using machine learning as a support tool to facilitate safe space for online discussions; for example, through detecting various types of negative online behaviors such as hate speech, online harassment, or cyberbullying. Typically, these efforts build upon sentiment analysis or spam detection in text. However, the toxicity of the language could be a strong indicator for the intensity of the negative behavior. In this paper, we study the topic of toxicity in online conversations by addressing the problems of subjectivity, bias, and ambiguity inherent in this task. We start with an analysis of the characteristics of subjective assessment tasks (e.g. relevance judgment, toxicity judgment, sentiment assessment, etc). Whether we perceive something as relevant or as toxic can be influenced by almost infinite amounts of prior or current context, e.g. culture, background, experiences, education, etc. We survey recent work that tries to understand this phenomenon, and we outline a number of open questions and challenges which shape the research perspectives in this multi-disciplinary field.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

ACM proceedings, crowdsourcing, subjectivity, toxicity

ACM Reference Format:
Lora Aroyo, Lucas Dixon, Olivia Redfield, Rachel Rosen, and Nithum Thain. 2019. Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317083

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317083

1 TOXICITY DETECTION

Online discussion forums currently provide a platform for many hazards, such as spreading fake news or expressing personal attacks in the form of online harassment or cyberbullying, typically identified by the use of toxic language in online discussions. Toxicity is a common term used to define negative online interactions in online spaces like Twitter[1], Wikipedia[2], and Google[3]. While the definition is often implied based on context, recent work has formalized one definition of toxicity as "comments that are rude, disrespectful or otherwise likely to make someone leave a discussion" (see, e.g. [21]).

A 2014 Pew Report[4] highlights that 73 percent of adult Internet users have seen someone being harassed online, and 40 percent have personally experienced it. The report outlines two categories of online harassment: (1) name-calling and embarrassment; this is typically considered less severe, however still annoying and often easier to ignore; (2) physical threats and harassment over a sustained period of time, such as stalking and sexual harassment; this category of harassment typically targets a smaller segment of the online population and involves more severe experiences. At the same time, the Wikimedia foundation found that 54 percent of those who had experienced online harassment expressed decreased online participation [47].

The prevalence and impact of this kind of harassment has triggered both the industrial and research communities in the recent years to focus on efficient computational models for online toxicity detection, specifically targeting detection of toxicity in online comments. However, steps in this direction are still in their infancy. Advances in machine learning triggered the development of a number of deep learning approaches displaying very promising performance so far. For example, [16] aims to discover toxic

[1]https://www.wired.com/story/twitter-let-users-hide-replies-fight-toxic-comments/
[2]https://arstechnica.com/information-technology/2017/02/one-third-of-personal-attacks-on-wikipedia-come-from-active-editors/
[3]https://www.perspectiveapi.com/
[4]http://www.pewinternet.org/2014/10/22/online-harassment/

comments in a large pool of documents provided by a current Kaggle competition regarding Wikipedia talk page edits. The authors compare CNNs against the traditional bag-of-words approach for text analysis combined with algorithms proven to be very effective in text classification. The results provide evidence that CNNs enhance toxic comment classification, reinforcing research interest in this direction. Another research initiative, founded by Jigsaw and Google, is Conversation AI[5]. The team is working on tools to help improve online conversation by researching how machine learning methods might help online conversations, what aspects of a conversation machine learning can understand, and what the risks and challenges of using machine learning to assist online conversations are. One specific area of focus in this context is the study of negative online behaviors, like toxic comments. The team has built a range of publicly available models served through the Perspective API[6], including a model that identifies whether a comment could be perceived as "toxic" according to the definition introduced above.

In this paper, we share our initial observations from surveying the current state of the art on detecting negative online behavior and we outline a number of open challenges related specifically to detecting toxicity:

- How does detecting toxicity relate to current approaches for detecting other forms of negative online behavior?
- How do bias and subjectivity impact toxicity assessment?
- How can we account for bias and subjectivity when designing crowdsourcing tasks for toxicity assessment?
- How can we ensure reproducible training and evaluation data for detecting and measuring toxicity in text?

The main contribution of this paper is providing input for a discussion on how we can move beyond the mere detection of sentiment or hate speech to a reliable assessment and ranking of toxicity by better understanding the influence of different personal, societal, and cultural perspectives. We focus the survey on related work on bias and subjectivity in human annotation and how these affect annotation of toxicity in text (section 1.1). We also dive into the current state of the art in detecting negative online behavior and identify the position of toxicity detection there (section 1.2). We position these in the larger context of crowdsourcing subjective assessments (section 1.3). Finally, we outline a number of challenges, open questions, and hypotheses related to measuring rater quality in such subjective tasks (section 2.1), evaluating the performance and reliability of different rating methods for collecting subjective assessments, especially the comparison of absolute rating vs. relative ranking (section 2.2), ensuring the reproducibility of results for toxicity annotation (section 2.3), learning how to rank from sample data (section 2.4), and identifying optimal task designs for subjective assessment tasks (section 2.5).

1.1 Bias and Subjectivity in Human Assessment

Psychologists Tversky and Kahnemann [42] first introduce the idea that judgments, especially subjective ones, can be prone to systematic human biases. Since then, this idea has been applied in various subjective assessment tasks. For example, [1] studies *preference bias* in the context of media recommender systems, showing the effect of anchoring on preference construction. [51] shows that listening tests designed to measure audio quality are vulnerable to systematic errors due to biases with respect to personal preferences, expectations, or mood. [12] exemplifies the "significant detrimental effects" of *cognitive biases* on the annotation quality in relevance judgment tasks, when such biases are not taken into consideration in the design of the task.

To ensure the quality of crowdsourced annotations, there has been a focus on reducing spam produced by bad actors, often by analyzing disagreement. However, spam is typically not the only source of noise in data– subjectivity and unintentional bias can cause disagreement among annotators (which ultimately may serve as a signal). For the purpose of this work, we consider two main categories of subjectivity:

(1) subjectivity inherent to the topic (ie. topics requiring judgments based on personal preference or experience, where two individuals may simply have differing opinions)
(2) subjectivity conditioned by ambiguity in the task (eg. ambiguity of guidelines, input items)

Both types of subjectivity can surface in a single annotation task. For example, in toxicity annotation, two annotators could have different levels of sensitivity to profanity, depending on their background. One person might consider the phrase "fucking awesome" to be toxic, while another would not; this type of subjectivity falls in category (1). Within the same task, they may encounter input items that are ambiguous or not covered by the guidelines that they have to make judgments on, which falls in category (2). It may be possible to mitigate (2) in task design, while (1) may surface regardless.

While the mainstream crowdsourcing approaches often treat disagreement as "noise" and aim their design to decrease and eliminate disagreement, there is an emerging community of researchers that indicate the importance of studying the disagreement and harnessing it to improve the quality of the crowdsourced data. [4] shows how the established practices in human annotation treat disagreement as "noise" and further [3] introduces the notion of "disagreement-based quality metrics", illustrating how the disagreement in crowdsourcing can be harnessed to achieve more representative and reliable annotation data. [24] proposes another approach for harnessing the disagreement through the *"crowd parting"* method to identify divergent, but valid, worker interpretations in crowdsourced tasks. [46] deals with a different approach to attend to the diversity of human annotators through a multi-dimensional measure of the annotators' ability. [38] demonstrates in the domain of semantic relatedness how the notion of a "universal gold standard" is highly problematic. The authors point out the importance of understanding these gold standard datasets from a human-centered point-of-view. Concretely, they empirically demonstrate that "people who belong to different cultural communities would provide different answers to a variety of knowledge-oriented tasks."

With respect to disagreement-aware task design, the CrowdVerge approach predicts which visual questions will have more disagreement and adds annotators to those questions in order to better capture the diversity of possible answers [19]. [8] explores limiting guidelines in favor of post-hoc label decisions, which reduces training time and surfaces potentially unanticipated ambiguity in

[5]https://conversationai.github.io/
[6]https://perspectiveapi.com

the task through disagreement. We explore open questions around task design further in section 2.5.

1.2 Detection of Negative Online Behaviours

As social media has become an unprecedentedly large and open space for information publishing and discussion, it is only natural that it would also attract forces that aim to exploit and misuse it to spread content that can be degrading, abusive, or otherwise harmful to people. To address this issue, research into automatic detection of different types of negative online behavior has emerged. While toxicity is a popular direction of study, it stands alongside other dimensions of negative online interaction. For example, [47] investigates personal attacks, [25] aggression, [7, 13, 45] hate speech, [29, 49] harassment, and [11] cyberbullying. [50] provides an in-depth overview of research in this area. There is also growing interest in understanding how these attributes relate to one another; for example, [20] investigates the connections between insulting, obscene, threatening, and hateful language and self-reported data on user behaviour.

As [47] points out, much of this work builds on existing machine learning approaches in fields like sentiment analysis [5, 35, 50, 52] and spam detection [37, 39]. However, one of the biggest challenges for detecting any of the forms of negative online behavior is the fact that all of them are influenced and defined by "the prevailing social norms, context, and individual and collective interpretation" [13]. Similarly, as seen in section 1.1, individual annotator differences make it difficult to identify such behavior consistently, and thus annotation tasks tend to result in high disagreement between annotators [27].

In addition to those challenges, it should be noted that even the definition of toxicity we discussed in Section 1 is highly context and community-dependent. What may be considered toxic on one platform or forum may be acceptable on another, due to the diverse nature of discourse in communities online.

1.3 Crowdsourcing Subjective Assessment

From the overview of research on bias and subjectivity (1.1) and how these influence and complicate the detection of negative online behaviours (1.2), we conclude that it is not trivial to ensure the reliability of human-annotated training data. It is critical to understand how each of these tasks inherently carry human bias and subjectivity, and to translate this to an adequate annotation task design.

One component that plays a central role in task design is the method for eliciting ratings. Absolute rating and comparative rating (or some combination of the two) are commonly leveraged. **Absolute rating** (often referred to as *judgment on a scale*) commonly asks annotators to rate the absolute value of some trait in the provided datum. There is extensive research ([30], [28], [9], [15], [18]) on different types of scales (e.g. Gunman scale, Likert scale at 7-point, 5-point, 3-point, or binary assessment, etc.) used to collect and measure user input, e.g. *nominal variables*: label a series of values; *ordinal scales*: order of values; *interval scales*: order of values and the ability to quantify the difference between each one; *ratio scales*: order, interval values and the ability to calculate ratios. **Comparative rating**, often referred to as *relative judgments*, frames the rating task as *a comparison*: asking raters to compare two items on a single characteristic; or in a *multiple comparison* setting, placing a number of items in order according to a single characteristic [23], [10]; [32], [14], [41], [48]. The research results of [17] support the hypothesis that *"humans may often be able to make more accurate ratings using comparative measures"* [26], [40], [34], [33].

We propose **hypothesis 1** that, especially in the case of toxicity assessments, agreement will differ between ratings produced by relative rating as opposed to an absolute scale. That is to say, we imagine it will be easier for raters to agree on a set of paired comparisons than it would be for them to agree on whether individual comments should be considered toxic.

Following the challenge of choosing the most appropriate assessment scale is that of generating a plausible rank from the collected judgments. In order to allow transferability of the results and their reuse within different ML systems, it is important to be able to convert results from one scale to another. In this process, however, it is critical to understand the loss of accuracy or meaning across scales. It could also be interesting to examine how raters' quality varies in these transformations, and also how we evaluate the accuracy of the transformation. When transforming either the absolute or relative ratings to linear rank, typically the distances between different points on the scale may not map uniformly.

2 CHALLENGES & OPEN QUESTIONS

Current models for capturing and understanding negative online behaviors, just like all machine learning models, still make errors. However, the challenges go beyond just improving the accuracy in detection. It is critical to provide more interactivity, transparency, and granularity for end users to be able to understand the range and diversity of the toxicity types (e.g. being able to select which types of toxicity they are interested in finding) and how they relate to specific context and cultures.

Thus, one of the grand challenges for such machine learning models is to understand the variety of human subjectivity and implicit (unintentional) bias. This is especially important when gathering training data for machine learning. One step towards solving this problem is the awareness that some data collection tasks carry intrinsic human subjectivity, and this needs to be understood in order to provide reliable training data. Research indicates that subjective assessment tasks are highly prone to disagreement between judges [43, 44]. This is typically caused by the ambiguity of the text and the inherent subjectivity of human raters. For example, Alonso and Mizzaro [2] found that the crowd relevance labels contradict the labels given by experts. However, individual worker judgments are still combined, without accounting for disagreement and ambiguity, by using majority vote or expectation maximization algorithms [31].

Below, we list a number of questions that deal with the challenges related to subjective task assessments.

2.1 Measuring rater quality in subjective tasks

As one of the goals of subjective task assessments is to gather the full range of human opinions, perspectives, and interpretations, inter-rater agreement is not an adequate measure for quality [4,

24]. We believe that metrics harnessing the diversity each rater contributes are needed in this context. However, the challenge is to distinguish between 'good' disagreement and 'bad' disagreement, which could be the result of spamming behavior [39]. We propose the following **hypothesis 2**: the distribution of raters disagreement will be indicative of raters perspectives on judging the toxicity of the text. In other words, the number of groups of ratings with similar counts can indicate the number of cultural or social perspectives with respect to the toxicity in the text. An additional **hypothesis 3** can be made - high disagreement in rater votes can indicate high ambiguity in the language used.

2.2 Evaluating different rating methods

To compare variants of the Revolt annotation approach with traditional crowdsourcing methods, [8] ran a series of annotation experiments, running each condition on eight different tasks with different datasets, ensuring no annotator saw more than one condition for a single dataset, and measuring the accuracy of the resulting annotations. Measuring effectiveness in subjective tasks can prove very challenging, as typically these tasks have no clear ground truth, or even if there is some notion of it, it can fluctuate over time. Thus, new metrics for measuring success and quality of the collected data need to be applied to capture or discriminate between different influences, e.g. time, culture, etc. A significant body of research focuses on how people rate differently on these scales, outlining the pros and cons of different scales (e.g. relative ratings vs. linear scores; Likert scales on different points compare; symmetric vs. asymmetric scales, etc) and how this impacts the reliability of the results.

Below we outline a number of observations and hypotheses with respect to the comparison of **relative rating vs linear scores**:

Severity of Toxicity: When rating a single comment, raters lack context about the full spectrum of toxicity among comments. As a result, a rater might rate a petty insult with the highest toxicity point on the scale without knowing that another comment in the pool contains more severe language, such as obscenities or threats. For example, when displayed independently, the comments "you are stupid" and "you should die" are both unambiguously toxic; consequently, they would likely be ranked the same in a single comment rating scheme, and models trained on this data would not contain any notion of severity. However, when displayed side by side, the difference between these comments becomes clearer, and the ratings are more likely to reflect that. Having raters understand this distinction is important because during an in-person conversation, the consequences of saying the threatening statement would be much harsher than those of saying the insulting one, and users often expect that models of online conversations will reflect these dynamics. We propose the following **hypothesis 4**: If a rating task requires rating a pair of comments relative to one another on a scale, then the overall comment rankings will more accurately reflect the severity of toxicity than the corresponding single comment task.

Keyword spotting: Certain words have strong connotations associated with them, and the perceived sentiment of the presence of certain language can have a strong effect on the perception of toxicity. However, the flexibility of language allows subtle changes in sentence structure to completely change the meaning of a phrase containing a specific word; examples include negation, quotation, prepositional phrases, and slang, among others. Consider the difference between the phrases "Go to hell" and "Hell yeah!", or between "Please don't hurt yourself" and "Please hurt yourself"; both sets of phrases contain similar words, but the toxicity of the comments differs completely. If seen in isolation, raters might only see the negative words in these phrases, but when asked to compare a pair, they may examine them more closely and find the distinction. We present the following **hypothesis 5**: If a relative rating task requires rating a pair of comments relative to one another on a scale, then there will be fewer errors due to "keyword spotting", which we define as a rater looking at words in isolation to make a judgment of toxicity, rather than the overall sentence.

Rating Variance: When presented with a comment with mild toxicity (e.g. a petty insult) in a pair of comments to compare, a rater might feel that the comment in question is considerably more toxic than the other comment (e.g. a friendly greeting), but given a separate comment to compare it to (e.g. a threat), the rater might feel that the comment in question is considerably less toxic than the other. This situation would yield two vastly different ratings on the Likert scale for two pairs containing the same comment, which would result in higher variance of ratings for that comment. Alternatively, when presented with a comment with extreme toxicity in a pair, the rater is likely to always say that the extremely toxic comment is much worse than any other comment it is compared to; this would yield lower variance among the ratings. With a single comment scale, all ratings for both the mildly toxic comment and the more extremely toxic comment are both likely to be on the higher end of the scale, since there is no comparison to be made; variance would be low in both cases. We present the following **hypothesis 6**: Relative rating variances of a comment considered across all pairs it could appear in should be higher than the absolute rating for the same comment, due to the range of contexts of the former (i.e. most comments appear alongside better and worse comments). The only comments with low relative rating variance should be those that are very obviously toxic (maxima in our ranking) or obviously not toxic, though the latter usually tends to be more ambiguous.

2.3 Ensuring reproducibility of results

It is important that human computation tasks be reproducible [36], but there are a number of bottlenecks related to reproducibility of results. One of the main challenges is the fact that individual annotation efforts are performed at a specific moment of time, while language, cultural context, and information evolve in the meantime. Each of these factors introduces a new perspective on the way text can be interpreted in a different period in time. Evolution of language online can be observed by studying social media data [22].

For our purposes, semantic change can be classified in two ways: **pejoration**, which is when a "word is used to express negatively loaded values not inherent in its historically original (or historically prior) meaning" and its opposite, **amelioration** [6]. Pejoration occurs more frequently and is therefore more relevant to the

classification of toxicity [6]; instances of pejoration of language introduce two potential causes of inconsistency into annotations of comments containing these newly toxic words:

(1) A rater more familiar with a newer, toxic context of a word could mistakenly rate a comment from years prior containing the word in a non-toxic context as toxic.
(2) A rater not familiar with the newer, toxic context of a phrase could mistakenly rate a recent toxic comment as non-toxic.

Corollaries to these cases would involve the same scenarios with amelioration of language.

Case (1) will inhibit reproducibility because the same experiment performed years later could yield false positives, contradicting the original negative annotations. Case (2) has a similar issue; if the experiment is repeated after the toxic phrase has become more well known, then the newer annotation will contradict the earlier, falsely negative annotation.

While it may inhibit reproducibility initially, case (2) will produce more accurate experimental results if the experiment continues to be repeated. Conversely, repetitions of case (1) will cause further divergence from the original results over time.

hypothesis 7: we hypothesize here that annotation efforts need to be performed continuously, so that annotated datasets can evolve with the latest linguistic, cultural and social contexts. Additionally, we hypothesize that capturing temporal snapshots of annotation efforts allows for a temporal analysis of the toxicity evolution with respect to severity and variance (as noted in section 2.2).

For instance, if we were to reproduce the crowdsourcing results from [47], using the same set of comments, same instructions, and same annotation platform, it would be important in our analysis to consider the temporal factor, where semantic change and cultural context may alter the results, as well as differences in annotator backgrounds.

2.4 Learning a rank from sample data

While collecting toxicity assessments at various scales, ultimately we aim at producing a ranked list of comments. The process of transforming the ratings at scale to a representative rank is not trivial for a number of reasons. We outline a few **open questions** that are important to consider in this context, e.g. what is an optimal amount of data to learn a rank; how well do probabilities correspond to severity; how well distributed are the values (linearization of the relative scores) vs the pre-selection methods based on scores from probabilities; how do sampling methods influence the optimal learning, and how to incorporate active learning in the process of collecting rating data?

2.5 Task Design for Subjective Tasks

In previous sections we hypothesized that disagreement is an indication of the differing annotator interpretations/perspectives and potential ambiguity in the text. In order to allow for this disagreement to be expressed and to enable its harnessing, the crowdsourcing task design should incorporate a number of elements. Thus, our **hypothesis 8** here is that in order to achieve an optimal design setting, a number of pilot experiments are needed with the following design considerations:

- what is an adequate *assessment scale* to deal with the variation and inconsistency of human judgments, both within and between subjects (see 2.2)
- what is an adequate *task template* that enables capturing the full spectrum of opinions, i.e. aiming at diversity in answers vs. optimizing for consensus (as indicated in sections 2.1 and 2.3)
- what is the optimal *number of raters* that will allow us to gather reliable and reproducible data; we hypothesize here that a higher number of raters will help to capture the full spectrum of diversity in opinions for tasks that are highly subjective or with highly ambiguous text.

ACKNOWLEDGMENTS

We acknowledge the contributions of our collaborators Chris Welty and Marie Pellat. We thank Lucy Vasserman and the workshop reviewers for their feedback.

REFERENCES

[1] Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. 2013. Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research* 24, 4 (2013), 956–975. https://doi.org/10.1287/isre.2013.0497 arXiv:https://doi.org/10.1287/isre.2013.0497
[2] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *IPM* 48, 6 (2012), 1053–1066.
[3] Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Journal of Human Computation* 1 (2014), 31–34. Issue 1. https://doi.org/10.15346/hc.v1i1.3
[4] Lora Aroyo and Chris Welty. 2015. Truth is a Lie: 7 Myths about Human Annotation. *AI Magazine* 36, 1 (2015).
[5] Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing State-of-the-Art Sentiment Models on State-of-the-Art Sentiment Datasets. *CoRR* abs/1709.04219 (2017). arXiv:1709.04219 http://arxiv.org/abs/1709.04219
[6] Paulina Borkowska and G. Kleparski. 2007. It befalls words to fall down: pejoration as a type of semantic change. *Studia Anglica Resoviensia* 47, 4 (2007), 33–50.
[7] Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making.. In *Proceedings of Internet, Policy and Politics Conference*. Oxford, United Kingdom.
[8] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI âĂŹ17*. ACM Press, 2334âĂŞ2346. https://doi.org/10.1145/3025453.3026044
[9] Dennis L Clason and Thomas J Dormody. 1994. Analyzing data measured by individual Likert-type items. *Journal of agricultural education* 35 (1994), 4.
[10] Herbet Aron David. 1963. The method of paired comparisons. (1963).
[11] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Trans. Interact. Intell. Syst.* 2, 3, Article 18 (Sept. 2012), 30 pages. https://doi.org/10.1145/2362394.2362400
[12] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 162–170. https://doi.org/10.1145/3159652.3159654
[13] Anirudh Srinivasan Adam Glynn Jacob Eisenstein Eshwar Chandrasekharan, Umashanthi Pavalanathan and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 1. ACM, New York, NY, USA, 22. https://doi.org/10.1145/3134666
[14] V.V. Fedorov. 1972. Theory of optimal experiments. *Elsevier* (1972).
[15] Gerhard H Fischer and Ivo W Molenaar. 2012. *Rasch models: Foundations, recent developments, and applications*. Springer Science & Business Media.
[16] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. 2018. Convolutional Neural Networks for Toxic Comment Classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence (SETN '18)*. ACM, New York, NY, USA, Article 35, 6 pages. https://doi.org/10.1145/3200947.3208069
[17] Richard D. Goffin and James M. Olson. 2011. Is It All Relative? Comparative Judgments and the Possible Improvement of Self-Ratings and Ratings of Others. *Perspectives on Psychological Science* 6, 1 (2011), 48–60. http://www.jstor.org/stable/41613423

[18] Joy Paul Guilford. 1954. Psychometric methods. (1954).

[19] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI âĂŹ17*. ACM Press, 3511âĂŞ3522. https://doi.org/10.1145/3025453.302578

[20] Toby Hopp, Chris J Vargo, Lucas Dixon, and Nithum Thain. 2018. Correlating Self-Report and Trace Data Measures of Incivility: A Proof of Concept. *Social Science Computer Review* (2018), 0894439318814241.

[21] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving GoogleâĂŹs Perspective API Built for Detecting Toxic Comments. *arXiv preprint arXiv:1702.08138* (2017).

[22] Noah A. Smith Jacob Eisenstein, Brendan O'Connor and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9, 11 (2014).

[23] Shaili Jain Anil K. Jain Jinfeng Yi, Rong Jin. 2013. Inferring UsersâĂŹ Preferences from Crowdsourced Pairwise Comparisons: A Matrix Completion Approach. *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing (HCOMP)* (01 2013), 207–215.

[24] Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. 1637âĂŞ1648.

[25] Joseph M Kayany. 1998. Contexts of uninhibited online behavior: Flaming in social newsgroups on Usenet. *Journal of the American Society for Information Science* 49, 12 (1998), 1135–1141.

[26] Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and BestâĂŞWorst Scaling. (2016).

[27] Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets Against Blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI'13)*. AAAI Press, 1621–1622. http://dl.acm.org/citation.cfm?id=2891460.2891697

[28] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).

[29] Melanie J Martin. 2002. Annotating flames in Usenet newsgroups: a corpus study. *For NSF Minority Institution Infrastructure Grant Site Visit to NMSU CS department* (2002).

[30] Michael S Matell and Jacob Jacoby. 1971. Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and psychological measurement* 31, 3 (1971), 657–674.

[31] Tyler McDonnell, Matthew Lease, Tamer Elsayad, and Mucahid Kutlu. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *HCOMP*. 10.

[32] Sir Moser, Claus and Graham Kalton. 1972. *Survey methods in social investigation* (2nd american ed ed.). New York : Basic Books. Previous ed.: 1958.

[33] Jack O'Neill, Sarah Delany, and Brian Mac Namee. 2018. From Rankings to Ratings: Rank Scoring via Active Learning. (10 2018).

[34] Jack O'Neill, Sarah Jane Delany, and Brian Mac Namee. 2017. Rating by Ranking: An Improved Scale for Judgement-Based Labels. In *IntRS@RecSys*.

[35] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2, 1-2 (Jan. 2008), 1–135. https://doi.org/10.1561/1500000011

[36] Praveen Paritosh. 2012. Human Computation Must Be Reproducible. In *CrowdSearch: WWW Workshop on Crowdsourcing Web Search*. 20–25.

[37] Vikas C. Raykar and Shipeng Yu. 2012. Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *J. Mach. Learn. Res.* 13 (Feb. 2012), 491–518. http://dl.acm.org/citation.cfm?id=2188385.2188401

[38] Rebecca Gold Benjamin Hillmann Matt Lesicko Samuel Naden Jesse Russell Zixiao (Ken) Wang Shilad Sen, Margaret E. Giesel and Brent Hecht. 2015. Turkers, Scholars, "Arafat" and "Peace": Cultural Communities and Algorithmic Gold Standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW âĂŹ15)*. ACM, New York, NY, USA, 826âĂŞ838. http://dx.doi.org/10.1145/2675133.2675285

[39] Nikita Spirin and Jiawei Han. 2012. Survey on Web Spam Detection: Principles and Algorithms. *SIGKDD Explor. Newsl.* 13, 2 (May 2012), 50–64. https://doi.org/10.1145/2207243.2207252

[40] Saif M. Mohammad Svetlana Kiritchenko. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. (2017).

[41] L. L. Thurstone. 1927. A Law of Comparative Judgment. (1927).

[42] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124 arXiv:http://science.sciencemag.org/content/185/4157/1124.full.pdf

[43] Ellen M Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *IPM* 36, 5 (2000), 697–716.

[44] Ellen M Voorhees. 2001. The philosophy of information retrieval evaluation. In *CLEF*, Vol. 1. Springer, 355–370.

[45] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 19–26.

[46] P. Welinder and P. Perona. 2010. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference*. 25âĂŞ32. http://dx.doi.org/10.1109/CVPRW.2010.5543189

[47] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on the World Wide Web*. Perth, 1391–1399.

[48] Bernardo A. Huberman Yarun Luon, Christina Aperjis. 2012. Rankr: A Mobile System for Crowdsourcing Opinions. (2012).

[49] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2 (2009), 1–7.

[50] Justine Zhang, Jonathan P Chang, Lucas Dixon, Yiqing Hua, Nithum Tahin, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1.

[51] Slawomir Zielinski, Francis Rumsey, and SÃ¿ren Bech. 2008. On Some Biases Encountered in Modern Audio Quality Listening Tests-A Review. *J. Audio Eng. Soc* 56, 6 (2008), 427–451. http://www.aes.org/e-lib/browse.cfm?elib=14393

[52] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Trans. Manage. Inf. Syst.* 9, 2, Article 5 (Aug. 2018), 29 pages. https://doi.org/10.1145/3185045