Joni Salminen, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J. Jansen (2021): The Problem of Majority Voting in Crowdsourcing with Binary Classes In: Proceedings of the 19th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.18420/ecscw2021_n12

Copyright 2021 held by Authors, DOI: 10.18420/ecscw2021_n12

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, contact the Authors.

The Problem of Majority Voting in Crowdsourcing with Binary Classes

Joni Salminen[1,2], Ahmed Mohamed Kamel[3], Soon-gyo Jung[1], Bernard J. Jansen[1]

[1]Qatar Computing Research Institute, Hamad Bin Khalifa University, [2]University of Turku, [3]University of Cairo
{jsalminen,sjung,bjansen}@hbku.edu.qa, ahmedm.kamel@pharma.cu.edu.eg

Abstract. When there are two classes, a majority vote can always be obtained with three labelers. Researchers can utilize this property to obtain a false sense of confidence in their ground truth labels. We demonstrate such a case with 3000 crowdsourced labels for an online hate dataset. Evaluating with percentage agreement, Gwet's AC1, and Krippendorff's alpha, results show that using more raters teases out the hidden nuances in raters' preferences. We show that full agreement among the raters monotonically decreases from three raters (28.4%) to nine raters (19.5%). Ten raters have a higher agreement than any other number of raters, which supports the idea of increasing the number of raters for subjective labeling tasks. Nevertheless, while beneficial, increasing the number of raters cannot be considered as a fundamental solution to the issue of agreement in subjective crowdsourcing tasks, as even with ten raters, there is a non-negligible number of ties (4.11%). We suggest having a small sample of the data labeled by five or more raters to evaluate the stability of agreement among the raters.

Introduction

Our argument focuses on the development of training sets for online hate detection (classification, scoring) models that are used in various computing systems. We argue that using binary classification with three raters can hide underlying disagreement among crowd raters. We suggest ways to tackle this issue.

Subjectivity in crowdsourced ratings is well-known (Alonso, 2011, 2015; Alonso et al., 2013, 2015; Alonso & Mizzaro, 2012; Aroyo et al., 2019; Salminen, Almerekhi, Dey, et al., 2018). Fundamentally, subjectivity means that individuals rate items differently based on personal beliefs, attitudes, worldviews, cultures, demographics, and other factors affecting their judgment (Alonso, 2015). Despite this, researchers dealing with subjective rating tasks, such as online hate/toxicity annotation, still use crowdsourced labels to construct training sets (Almerekhi et al., 2019, 2020; Davidson et al., 2019, 2017; Fortuna, 2017; Vidgen & Derczynski, 2020) for machine learning (ML). Crowdsourcing, in general, refers to using an anonymous pool of users to carry out human intelligence tasks (HITs) (Kittur, Chi, et al., 2009; Kittur et al., 2008, 2013; Kittur, Lee, et al., 2009; Yu et al., 2016).

In this research, we investigate the particular case of majority voting in a binary labeling task when using crowdsourced ratings. Binary labeling refers to a task where the raters have two options (e.g., yes/no, positive/negative). Majority voting refers to using the "winning" class as the final ground truth label. For example, if two raters say "yes" and one says "no," then the final label is yes (2/3). Similarly, if there are five raters, then the class obtaining three or more votes will be the final label, and so on. With an odd number of raters, the binary classification will always have a majority label when using majority voting.

Our research goal is to evaluate if majority voting is a justified strategy for binary classification when the task has a non-negligible degree of subjectivity (i.e., room for interpretation). To investigate this matter, we collect 3,000 ratings on 300 social media comments from a crowdsourcing platform and investigate how the dynamics of inter-rater agreement evolve when varying the number of raters.

We chose hate detection as the illustrative context for three reasons: (1) Prevalence of hate in online social media, (2) hate labeling has known issues of subjectivity and interpretation (Fortuna & Nunes, 2018; MacAvaney et al., 2019; Modha et al., 2020; Sood et al., 2012), and (3) there are several examples of studies (Davidson et al., 2017; Ibrohim & Budi, 2019; Magdy et al., 2015) applying the majority rule in crowdsourced labels to achieve ground-truth labels in this space.

Online hate detection is a growing field of research (see reviews in (Fortuna & Nunes, 2018; Waqas et al., 2019)) with broad cross-disciplinary interest among scholars from different communities, including HCI (Türkay et al., 2020). Typically, hate detection involves ML models with crowdsourced ratings as training data (Davidson et al., 2017; Mohan et al., 2017; Mondal et al., 2017; Salminen, Almerekhi, Milenković, et al., 2018; Waseem, 2016). A prominent

example is Perspective API (Alphabet, 2018), a tool by Jigsaw to score online comments for toxicity and hate. However, dataset quality is considered one of the most pressing challenges in online hate detection (MacAvaney et al., 2019; Modha et al., 2020; Vidgen & Derczynski, 2020). Our starting point for this study is that more research is needed into understanding the subjective nature of online hate and how this affects the training set creation process.

Related Literature

In HCI, crowdsourcing has been applied for various tasks, including taxonomy creation (Chilton et al., 2013), user studies (Kittur et al., 2008) such as graphical perception (Heer & Bostock, 2010), user interface performance (Komarov et al., 2013), accessibility (Hara et al., 2013), as well as generation of creative design outputs (Willett et al., 2012). Beyond HCI, social computing studies apply crowdsourcing to generate training samples for ML models (Davidson et al., 2017; Huang et al., 2014; Kocabey et al., 2018; Weber & Mejova, 2016).

ML techniques and crowdsourcing are a powerful combination for learning about online users. Nevertheless, the quality of the obtained annotations does not always lead to questions of dataset reliability (Alonso et al., 2013). Researchers tend to measure the inter-rater agreement as a proxy for quality (Alonso et al., 2015) to avoid such quality issues, with metrics such as Cohen's kappa (Cohen, 1960), Krippendorff's alpha (Krippendorff, 1980), Gwet's AC1 (Gwet, 2008), and others (Banerjee et al., 1999). When these metrics show a lack of agreement, several potential explanations arise. For example, guidance and instructions given to the raters may be inadequate or unclear (Pitkänen & Salminen, 2013), there may be fraudulent raters or bots (Peng et al., 2014) or there may be a sincere lack of attention (Alonso, 2015). A particular problem is *inherent subjectivity* (Salminen, Almerekhi, Dey, et al., 2018), meaning that the task actually has *no right or wrong answer*. To solve the issue of inherent subjectivity, researchers can deploy an odd number of raters and choose the last rater as a tiebreaker to assign the final label for the classified sample (Duwairi et al., 2014; Ibrohim & Budi, 2019; Magdy et al., 2015; Trieu et al., 2017; Volkova & Yarowsky, 2014).

Hate labeling is an example of a subjective labeling task. This is because individuals' opinions of what constitutes a hateful comment might differ despite the fact that a commonly accepted definition is provided (Alonso, 2015; Salminen et al., 2019; Salminen, Veronesi, et al., 2018). In (Salminen et al., 2019), the researchers analyzed 5,665 crowd ratings on 1,133 social media comments. The results indicated that individuals tend to agree on the extremes of a hate rating scale more than in the middle. The agreement was higher for comments that were, on average, considered less hateful and lower on comments that were generally rated as moderately hateful. The researchers suggest that this behavior helps reach an agreement on extreme cases (very hateful/not hateful at all) faster and more cost-

efficiently than obtaining an agreement on gray-area cases. In (Salminen, Veronesi, et al., 2018), the researchers collected 18,125 ratings from crowd workers in 50 countries, analyzing the effect of the country on the given hate scores. Even though geographic patterns were found, the conclusion is that hate ratings vary more by the individual raters than by countries.

There are many other subjective labeling tasks beyond hate detection. Examples include sentiment analysis (Cambria, 2016), peer-nominated personality ratings (Celli, 2011; Celli & Rossi, 2012), or virtually any topic dealing with opinions, attitudes, and preferences. The key distinction between subjective labeling tasks for ML applications of crowdsourcing is that they are conducted for the purpose of building a training set. This purpose tends to come with the explicit requirement of ground truth (Weber, 2015)—i.e., an assumption that the items have one true value. This is not the case for surveys, for which it is generally accepted (and expected) that the respondents' answers vary. In contrast, variation is a *problem* in a labeling task whose purpose is training set creation.

Method

Data Collection

We randomly sampled 300 comments from a previously published online hate dataset with known ground truth values (Salminen, Almerekhi, Milenković, et al., 2018). Half (n=150) of the comments are marked as hateful in the dataset, the other half as neutral. The crowd raters were recruited using the Appen platform (formerly known as CrowdFlower). The raters were presented with a simple binary decision task: "Is this comment hateful?" (Yes/No).

The raters were provided the following definition of hatefulness, similar to the definition applied by the dataset source (Salminen, Almerekhi, Milenković, et al., 2018): "*A hateful comment is rude, disrespectful or otherwise likely to make someone leave a discussion*."

We chose to have each sample labeled by ten raters. This choice is arbitrary, and we could also have aimed at having twenty or thirty raters as well. However, since this would have doubled or tripled the cost of acquiring data, we decided to choose ten raters. Overall, collecting 300 × 10 = 3,000 ratings was considered as enough data to examine the majority voting problem in crowdsourced labels. Also, according to the crowdsourcing platform's suggestions, eight test questions were created. For these test questions, we selected cases that were (according to our interpretation) clearly in either category (see Table 1). This decision is also in line with the best practices of the platform to make sure the test questions are as unambiguous as possible in order to avoid false rejection of workers, which will also result in extra cost for the crowdsourcing task.

Table 1. Test questions for the quality control of the crowdsourcing task. “Correct rate” indicates the share of correct crowd evaluations in the total. The observed correct rates are well aligned with the ground truth classes.

Comment	Class	Correct rate
Usain Bolt my hero has donated 10 million dollars to Haiti, all governments should hang their head in shame. A 100 metres runner did more than you. Bless up king Bolt	Non-hateful	92%
Chicago pizza is by far the best	Non-hateful	96%
Land ownership and the quiet title legal process in Hawaii are super complicated. For a better understanding of how it works, check out this resource from the Native Hawaiian Legal Corporation: http://www.nhlchi.org/resources/brochures/13#what-is-a-quiet-title-action	Non-hateful	95%
Fuck Obama and his Muslim apologism	Hateful	99%
he is not a fucking refugee he is a economic migrant fuck aj+	Hateful	100%
Stick your PROPAGANDA right up your FUCKING MUSLIM ASS!!!!!!!!!!!!!!!!!!!!	Hateful	98%
Fucking bullshit regressive channel. Aj+ fucking Muslim terrorist apologists. They won’t be singing in front of them when one of the Syrians or their decedents blow them up. I wonder what these Syrians think of Jews. I doubt they have a high opinion of them because their anti Semitic because their muslim	Hateful	99%
even stupid george bush knows its called quds force and not al quds, so fuck this channel and this whore.. this is a fucking jew channel, calling quds force al quds make it sound like al qaeda and al shabab and al nusra and those arab terrorist monkeys this is a persian force	Hateful	99%

Other quality control settings are also in line with the platform’s recommendations:

1. **Minimum Time per Page = 10 Seconds (Default).** This is the minimum time raters are required to complete a page of annotations. If less time is spent, the rater will be removed from the task.
2. **Disable Google Translate For Contributors = Enabled.** When enabled, this option disables Google Translate for raters using the Chrome browser to ensure that context and meaning are not lost in translation.
3. **Max Judgments per Contributor = Empty (Default).** This setting limits the maximum number of ratings that a rater can provide for the task. By default, the maximum ratings a rater can submit is limited by the number of test questions in the task. (In our case, eight test questions.)
4. **Quality Level** = 2 (“higher quality: a smaller group of more experienced contributors with a higher accuracy”).

The compensation for the workers was set at **rows per page = 5 (default)** and **price per page = USD 35 cents per page (default)**. These settings resulted in the **price per judgment = USD 7 cents (default)**. The parameters were set based on the belief that the platform’s defaults respect the minimum pay guidelines for crowdsourcing (Vaughan, 2017). The total cost for data collection was $316.26.

In other words, apart from the translation prevention and the increase of the quality level from default 1 to the higher level of 2, the other options were default.

We set the geographic targeting to the United States to gain some control over the cultural factors in the interpretation of hateful social media comments (Mubarak et al., 2017; Mubarak & Darwish, 2019).

Data Cleaning

A total of 300 social media comments were rated for hatefulness using crowdsourced ratings on a binary scale (yes/no). Each comment was rated by ten raters. The eight test questions that were rated by more than ten raters were excluded from the analysis for parsimony. Thus, the analysis comprised 292 comments with a total of 2,920 ratings. The ratings were sorted chronologically by creation date within each comment prior to the analysis.

Statistical Analysis

The statistical analysis was performed using the R software (v. 3.6.3). Counts and percentages were used to summarize the variables. The inter-rater reliability was assessed by using three measures: (1) percentage agreement, (2) Gwet's AC1 (Gwet, 2008), and (3) Krippendorff's alpha (K alpha). Using multiple agreement measures is advisable to ensure the consistency of the results (Cicchetti & Feinstein, 1990) by mitigating the impact of the shortcomings of any given metric on the overall findings. The AC1 and K alpha are chance-corrected agreement measures.

The K alpha measure can be used for nominal and ordinal outcomes and can take a value between 0 (perfect disagreement) and 1 (perfect agreement). It can also accommodate missing data, although in this case, we had none. The K alpha corrects the expected agreement by chance and can acquire lower values with high values of percentage agreement (Krippendorff, 1980).

The AC1 can be used when the expected agreement due to chance is high, which inversely affects the calculation of K alpha (Gwet, 2008). AC1 was developed as an alternative method in the presence of high expected agreement by chance, as it does not assume independence between raters. AC1 also supports categorical, ordinal, interval and ratio types of data and supports missing values.

Results

Number of ties based on the number of raters

We calculate the number of ties to understand how much using majority voting would affect the final labels. A tie is a situation where an equal number of raters think the comment is hateful and non-hateful (e.g., out of six raters, three choosing "yes" and three choosing "no" constitutes a tie).

Table 2 shows two important results. First, *there are no tie ratings, ever, when using an odd number of raters*. That is, even if there is an underlying tendency of disagreement among the raters, this can be obfuscated by choosing an odd number of raters and their majority decision on a given item.

Table 2. Ties for ratings with even raters

Raters	Number of ratings
2	79 (27%)
3	0%
4	37 (12.7%)
5	0%
6	26 (8.9%)
7	0%
8	16 (5.48%)
9	0%
10	12 (4.11%)

Second, *the proportion of comments with ties decreases with the increase in the number of raters*. Ties were observed for 79 (27%) and 37 (12.7%) comments when the ratings from the first two and four raters were used for the analysis, respectively. The number decreased to 26 (8.9%) when the ratings from six raters were used and further decreased to 16 (5.48%) when eight raters were used. The number of ties was lowest when all ten raters were used for the analysis (4.11%).

This finding can be interpreted, in a certain sense, as convergence to a consensus opinion on the "true" ratings of the items (see Figure 1). However, it is notable that *even with ten raters, there is a non-negligible number of ties*.

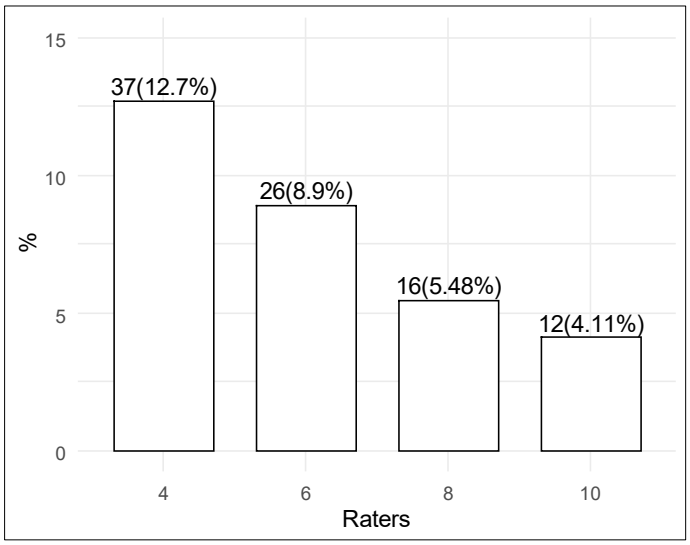


Figure 1: Ties based on the number of raters. The decreasing number indicates convergence to the "true" values. Yet, even with ten raters, some cases have equal support of being "yes" or "no."

Agreement on the Hatefulness of Comments

We next analyze the structure of the majority vote among the raters. This analysis is based on the majority rule, i.e., the comment was deemed hateful if more than 50% of the raters found it hateful. The proportion of comments that fit this criterion was calculated based on the number of raters. The analysis was performed for comments with an odd number of raters. Results in Table 3 show three interesting findings: First, the frequency of hateful comments corresponds well with the expected frequency (the ground truth had 150 comments labeled hateful, and the raters found 145-148 hateful comments). Second, the frequency of hateful comments remains stable from three to nine raters (at around 50%). Third, a full agreement among the raters (i.e., all of the raters agreeing that a comment is hateful) monotonically decreases from three raters (28.4%) to nine raters (19.5%).

Table 3. Structure of the majority vote on hateful comments.

Hateful comments (> 50%)		**Raters who found the comment hateful**							
	Raters	2	3	4	5	6	7	8	9
147 (50.3%)	3	64 (21.9%)	83 (28.4%)						
147 (50.3%)	5		30 (10.3%)	47 (16.1%)	70 (24%)				
148 (50.7%)	7			21 (7.19%)	26 (8.9%)	39 (13.4%)	62 (21.2%)		
145 (49.7%)	9				14 (4.79%)	17 (5.82%)	23 (7.88%)	34 (11.6%)	57 (19.5%)

According to the majority rule, with three raters, 100% of the comments have at least 66.7% agreement (2/3). However, what is the proportion of comments with five, seven, or nine raters with at least 66.7% agreement? Mathematically, the "worst" case for a given class to win decreases as the number of raters increases. For five, it is 3/5 = 60%; for seven, it is 4/7 = 57.1%; for nine, it is 5/9 = 55.6%. Our data shows that when the same comments are evaluated by five raters, 40.1% (n=117) of the comments have at least 66.7% agreement. With seven raters, 43.5% (n=127) have at least 66.7% agreement, and for nine raters, the value is 44.9% (n=131). These results imply that *using more raters helps understand the subjectivity of the task by teasing out differences in the agreement structure*.

Inter-rater Reliability

Our third analysis focuses on the agreement among the raters in all instances. The results in Table 4 show that the percentage agreement rate varies from 71.5% to 75.1%. The K alpha and AC1 measures were significantly different from zero, irrespective of the number of raters, as shown by the 95% confidence intervals

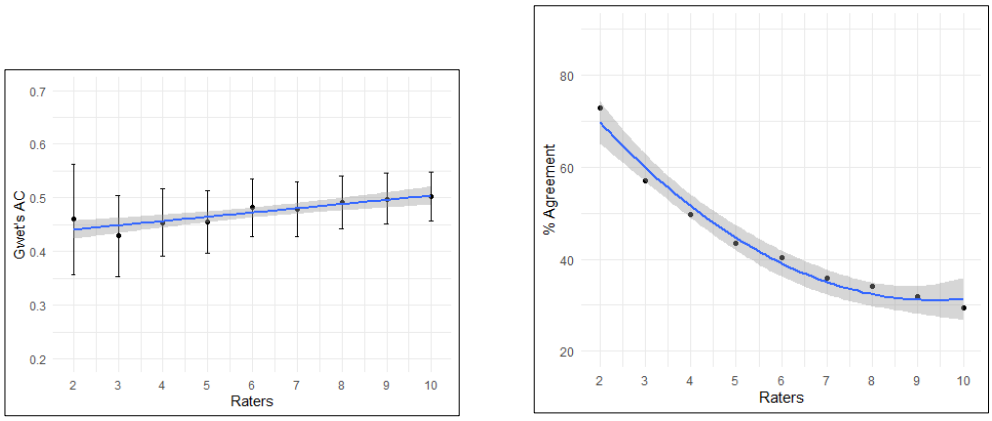
above 1. Interestingly, the agreement remains fairly stable throughout the increase in the number of raters. Ten raters have a higher agreement rate than any other number of raters, supporting the increase in the number of raters.

Table 4. Inter-rater reliability scores (95% confidence intervals in parentheses). The metrics show consistent results.

Raters	% agreement	K alpha	Gwet's AC
2	72.9%	0.459 (0.356, 0.561)	0.46 (0.357, 0.563)
3	71.5%	0.43 (0.354, 0.506)	0.429 (0.353, 0.505)
4	72.7%	0.455 (0.391, 0.518)	0.454 (0.391, 0.518)
5	72.7%	0.455 (0.397, 0.512)	0.455 (0.397, 0.513)
6	74.1%	0.481 (0.428, 0.534)	0.482 (0.428, 0.535)
7	73.9%	0.478 (0.427, 0.528)	0.479 (0.428, 0.53)
8	74.6%	0.491 (0.443, 0.539)	0.492 (0.443, 0.54)
9	74.9%	0.498 (0.451, 0.545)	0.498 (0.451, 0.546)
10	75.1%	0.502 (0.457, 0.548)	0.502 (0.457, 0.548)

The results in Figure 2a show that the 95% confidence intervals are overlapping, although the values tended to be slightly higher with the increase in the number of raters. Regression analysis was used to assess whether a statistically significant linear trend existed in the relation between the number of raters and AC1. Data points were weighted using the inverse of the standard error, so data points with a higher standard error (less confidence) had lower weight in the regression analysis. The results indicate a statistically significant positive linear trend ($B = 0.008$, $P < 0.001$). This indicates that increasing the number of raters is associated with a modest but significant increase in AC1.

Finally, a linear regression analysis shows a statistically significant quadratic trend (see Figure 2b) in the relation between the number of raters and the perfect agreement rate ($P < 0.001$) with a strong initial decline in the proportion of raters who were in perfect agreement and a slightly less strong association at later stages (after adding a 6th or 7th rater).



(a) **(b)**

Figure 2: **(a)** Gwet's AC1 based on the number of raters (the vertical lines represent the 95% confidence interval, and the horizontal line represents the regression line); **(b)** Perfect agreement based on the number of raters. The smoothed line represents the negative quadratic trend when increasing the number of raters.

Discussion and Practical Implications

In summary, using the majority vote tactic with three raters and binary classification is not recommended as the only option for building ML datasets, as this can cloud the subjectivity of the task and give a false sense of dataset validity. Even though the study only tested an online hate rating task, similar results are to be expected for other subjective rating tasks.

The results also imply that, while beneficial, increasing the number of raters cannot be considered the fundamental solution to the issue of agreement in subjective crowdsourcing tasks. Subjectivity can be so strongly ingrained in the data that no number of raters results in perfect agreement.

If uncertain, researchers can probe the subjectivity of their task by annotating a small sample of data with a large number of raters and observe how agreement and majority vote tendencies evolve. Another option is to sway from the requirement of one true label for every item in the ground truth. Instead, researchers can investigate the use of empirical distributions, as done in (Wulczyn et al., 2017). Essentially, whereas the one-true-label approach requires the predicted value to be either 1 (hateful) or 0 (non-hateful), the empirical description contains a tuple of values (e.g., [0.7, 0.3]). Hence, there is more information on the distribution of preferences. Depending on the number of classes and the readiness of the applied ML algorithm, empirical distributions can have a varying number of elements.

Previous research suggests that extremely hateful or non-hateful comments reach a consensus faster than comments in the mid-range (Salminen et al., 2019). Yet, we are not aware of any annotation schema that would leverage this property.

Some platforms such as Appen offer "dynamic judgments," a feature that collects *more* ratings for samples that struggle to reach consensus. However, what perhaps would be needed is the *exclusion* of such samples. If a sample is inherently subjective, collecting more ratings will not help resolve disagreements. Alternatively, these grey area comments could be labeled as such – e.g., apply label "indecisive" and use that as a third category for training the hate classifier.

While this analysis focused on hate detection datasets as the context, our findings apply to other training set creation tasks, e.g., those in the realm of NLP and sentiment analysis (Cambria, 2016; Celli, 2011; Celli & Rossi, 2012), as these fields generally face the same systematic issue of subjectivity.

Finally, we would like to point out that the are some general limitations when relying on crowd work for research purposes. For example, the lack of subject-matter expertise may be harmful to ML outcomes when the training data annotation would require specific domain knowledge (Alonso, 2015; Alonso et al., 2013). When recruiting crowd workers, this issue can partially be addressed by including training as a part of the annotation process (e.g., by using test questions that clarify where the crowd worker made a mistake), but this is not possible when the required level of expertise exceeds what can reasonably be trained in a short amount of time. Overall, researchers may benefit from expanding their views of how to design a crowdsourcing task, including questions about whether the ground truth unfolds as a result of a planned process, series of clarifications and redefinitions, or as a succession of surprises and repairs (Muller et al., 2021).

Conclusion

When tasks are subjective, using crowdsourced majority voting with three raters can hide real disagreements. Our results show that the rate of perfect agreement decreases with the increase in the number of raters. Researchers can label a small sample of their data with more than three raters (e.g., 5, 10) to validate the stability of their ground truth labels before conducting further analyses.

References

Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2020). Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions. Proceedings of The Web Conference 2020, 3033–3040. https://doi.org/10.1145/3366423.3380074

Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2019). Detecting Toxicity Triggers in Online Discussions. *Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT'19)*, 291–292. https://doi.org/10.1145/3342220.3344933

Alonso, O. (2011). Crowdsourcing for Information Retrieval Experimentation and Evaluation. In P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, & M. de Rijke (Eds.), *Multilingual and Multimodal Information Access Evaluation* (Vol. 6941, pp. 2–2). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23708-9_2

Alonso, O. (2015). Practical Lessons for Gathering Quality Labels at Scale. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1089–1092. https://doi.org/10.1145/2766462.2776778

Alonso, O., Marshall, C. C., & Najork, M. (2015). Debugging a Crowdsourced Task with Low Inter-Rater Agreement. *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 101–110. https://doi.org/10.1145/2756406.2757741

Alonso, O., Marshall, C. C., & Najork, M. A. (2013, November 3). A Human-Centered Framework for Ensuring Reliability on Crowdsourced Labeling Tasks. *First AAAI Conference on Human Computation and Crowdsourcing*. First AAAI Conference on Human Computation and Crowdsourcing. https://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7487

Alonso, O., & Mizzaro, S. (2012). Using crowdsourcing for TREC relevance assessment. *Information Processing & Management*, *48*(6), 1053–1066. https://doi.org/10.1016/j.ipm.2012.01.004

Alphabet. (2018). *Perspective API*. https://www.perspectiveapi.com/#/

Aroyo, L., Dixon, L., Thain, N., Redfield, O., & Rosen, R. (2019). Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. *Companion Proceedings of The 2019 World Wide Web Conference*, 1100–1105. https://doi.org/10.1145/3308560.3317083

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*(1), 3–23. https://doi.org/10.2307/3315487

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, *31*(2), 102–107.

Celli, F. (2011). Mining user personality in twitter. *Language, Interaction and Computation CLIC*.

Celli, F., & Rossi, L. (2012). The role of emotional stability in Twitter conversations. *Proceedings of the Workshop on Semantic Analysis in Social Media*, 10–17.

Chilton, L. B., Little, G., Edge, D., Weld, D. S., & Landay, J. A. (2013). Cascade: Crowdsourcing taxonomy creation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1999–2008.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558. https://doi.org/10.1016/0895-4356(90)90159-M

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *Proceedings of the Third Workshop on Abusive Language Online*, 25–35. https://doi.org/10.18653/v1/W19-3504

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of Eleventh International AAAI Conference on Web and Social Media*, 512–515.

Duwairi, R. M., Marji, R., Sha'ban, N., & Rushaidat, S. (2014). Sentiment analysis in arabic tweets. *2014 5th International Conference on Information and Communication Systems (ICICS)*, 1–6.

Fortuna, P. (2017). Automatic detection of hate speech in text: An overview of the topic and dataset annotation with hierarchical classes [Master's thesis]. Faculdade De Engenharia Da Universidade Do Porto.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1–30.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48. https://doi.org/10.1348/000711006X126600

Hara, K., Le, V., & Froehlich, J. (2013). Combining crowdsourcing and google street view to identify street-level accessibility problems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 631–640.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 203–212.

Huang, W., Weber, I., & Vieweg, S. (2014). Inferring nationalities of Twitter users and studying inter-national linking. *Proceedings of the 25th ACM Conference on Hypertext and Social Media - HT '14*, 237–242. https://doi.org/10.1145/2631775.2631825

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian twitter. *Proceedings of the Third Workshop on Abusive Language Online*, 46–57.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, 453–456. https://doi.org/10.1145/1357054.1357127

Kittur, A., Chi, E. H., & Suh, B. (2009). What's in Wikipedia?: Mapping Topics and Conflict Using Socially Annotated Category Structure. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1509–1512. https://doi.org/10.1145/1518701.1518930

Kittur, A., Lee, B., & Kraut, R. E. (2009). Coordination in collective intelligence: The role of team structure and task interdependence. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1495–1504.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., & Horton, J. (2013). The future of crowd work. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 1301–1318. http://dl.acm.org/citation.cfm?id=2441923

Kocabey, E., Ofli, F., Marin, J., Torralba, A., & Weber, I. (2018). Using computer vision to study the effects of BMI on online popularity and weight-based homophily. *International Conference on Social Informatics*, 129–138.

Komarov, S., Reinecke, K., & Gajos, K. Z. (2013). Crowdsourcing performance evaluations of user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 207–216. https://doi.org/10.1145/2470654.2470684

Krippendorff, K. (1980). Validity in Content Analysis. *Computerstrategien Für Die Kommunikationsanalyse*, 69–112.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*, *14*(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Magdy, W., Darwish, K., & Abokhodair, N. (2015). Quantifying public response towards Islam on Twitter after Paris attacks. *ArXiv Preprint ArXiv:1512.04570*.

Modha, S., Mandl, T., Majumder, P., & Patel, D. (2020). Tracking Hate in Social Media: Evaluation, Challenges and Approaches. *SN Computer Science*, *1*, 1–16.

Mohan, S., Guha, A., Harris, M., Popowich, F., Schuster, A., & Priebe, C. (2017). The Impact of Toxic Language on the Health of Reddit Communities. *Proceedings of the Canadian Conference on Artificial Intelligence*, 51–56. https://doi.org/10.1007/978-3-319-57351-9_6

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A Measurement Study of Hate Speech in Social Media. *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 85–94.

Mubarak, H., & Darwish, K. (2019). Arabic offensive language classification on twitter. *International Conference on Social Informatics*, 269–276.

Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive language detection on Arabic social media. *Proceedings of the First Workshop on Abusive Language Online*, 52–56.

Muller, M., Wolf, C., Andres, J., Desmond, M., Joshi, N. N., Ashktorab, Z., Sharma, A., Brimijoin, K., Pan, Q., & Duesterwald, E. (2021). Designing Ground Truth and the Social Life of Labels. *Proceedings of ACM Human Factors in Computing Systems (CHI'21)*.

Peng, L., Xiao-yang, Y., Yang, L., & Ting-ting, Z. (2014). Crowdsourcing fraud detection algorithm based on Ebbinghaus forgetting curve. *International Journal of Security and Its Applications*, *8*(1), 283–290.

Pitkänen, L., & Salminen, J. (2013, November). Managing the Crowd: A Study on Videography Application. *In Proceedings of Applied Business and Entrepreneurship Association International (ABEAI)*.

Salminen, J., Almerekhi, H., Dey, P., & Jansen, B. J. (2018, October 15). Inter-rater agreement for social computing studies. *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018)*. The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018), Valencia, Spain.

Salminen, J., Almerekhi, H., Kamel, A. M., Jung, S., & Jansen, B. J. (2019). Online Hate Ratings Vary by Extremes: A Statistical Analysis. *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 213–217. https://doi.org/10.1145/3295750.3298954

Salminen, J., Almerekhi, H., Milenković, M., Jung, S., An, J., Kwak, H., & Jansen, B. J. (2018, June 25). Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. *Proceedings of The International AAAI Conference on Web and Social Media (ICWSM 2018)*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17885

Salminen, J., Veronesi, F., Almerekhi, H., Jung, S., & Jansen, B. J. (2018, October 15). Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings. *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018)*. The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018), Valencia, Spain.

Sood, S., Antin, J., & Churchill, E. (2012). Profanity Use in Online Communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1481–1490.

Trieu, L. Q., Tran, H. Q., & Tran, M.-T. (2017). News classification from social media using twitter-based doc2vec model and automatic query expansion. *Proceedings of the Eighth International Symposium on Information and Communication Technology*, 460–467.

Türkay, S., Formosa, J., Adinolf, S., Cuthbert, R., & Altizer, R. (2020). See No Evil, Hear No Evil, Speak No Evil: How Collegiate Players Define, Experience and Cope with Toxicity. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Vaughan, J. W. (2017). Making better use of the crowd: How crowdsourcing can advance machine learning research. *The Journal of Machine Learning Research*, *18*(1), 7026–7071.

Vidgen, B., & Derczynski, L. (2020). Directions in Abusive Language Training Data: Garbage In, Garbage Out. *ArXiv Preprint ArXiv:2004.01670*.

Volkova, S., & Yarowsky, D. (2014). Improving gender prediction of social media users via weighted annotator rationales. *NIPS 2014 Workshop on Personalization*.

Waqas, A., Salminen, J., Jung, S., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLOS ONE*, *14*(9), e0222194. https://doi.org/10.1371/journal.pone.0222194

Waseem, Z. (2016). Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter. *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142.

Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower*, *36*(1), 13–25.

Weber, I., & Mejova, Y. (2016). Crowdsourcing health labels: Inferring body weight from profile pictures. *Proceedings of the 6th International Conference on Digital Health Conference*, 105–109.

Willett, W., Heer, J., & Agrawala, M. (2012). Strategies for crowdsourcing social data analysis. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 227–236. https://doi.org/10.1145/2207676.2207709

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399. https://doi.org/10.1145/3038912.3052591

Yu, L., Kittur, A., & Kraut, R. E. (2016). Encouraging "Outside- The- Box" Thinking in Crowd Innovation Through Identifying Domains of Expertise. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1214–1222. https://doi.org/10.1145/2818048.2820025