Inter-rater Agreement for Social Computing Studies

Joni O. Salminen[1], Hind A. Al-Merekhi[2], Partha Dey[3] and Bernard J. Jansen[1]
[1] Qatar Computing Research Institute
Hamad Bin Khalifa University, Doha, Qatar
Email: jsalminen@hbku.edu.qa, jjansen@acm.org
[2] The College of Science and Engineering
Hamad Bin Khalifa University, Doha, Qatar
halmerekhi@qf.org.qa
[3] Department of Mechanical Engineering, Academy of Technology, Adisaptagram, India
pdey.bit@gmail.com

***Abstract*—Different agreement scores are widely used in social computing studies to evaluate the reliability of crowdsourced ratings. In this research, we argue that the concept of agreement is problematic for many rating tasks in computational social science because they are characterized by subjectivity. We demonstrate this claim by analyzing four social computing datasets that are rated by crowd workers, showing that the agreement ratings are low despite deploying proper instructions and platform settings. Findings indicate that the more subjective the rating task, the lower the agreement, suggesting that tasks differ by their inherent subjectivity and that measuring the agreement of social computing tasks might not be the optimal way to ensure data quality. When creating subjective tasks, the use of agreement metrics potentially gives a false picture of the consistency of crowd workers, as they over-simplify the reality of obtaining quality labels. We also provide empirical evidence on the stability of crowd ratings with a different number of raters, items, and categories, finding that the reliability scores are most sensitive to the number categories, somewhat less sensitive to the number of raters, and the least sensitive to the number of items. Our findings have implications for computational social scientists using crowdsourcing for data collection.**

***Index Terms*—crowdsourcing, crowd evaluations, crowd ratings, inter-rater reliability, social computing**

I. Introduction

Computational social scientists are frequently using crowdsourcing to label training data for developing machine learning models, often for classification purposes [1]. Training data refers to samples annotated by humans (e.g., crowd workers or experts) that are used to teach an algorithm to predict labels of unseen data [2]. For example, a social media comment "Pizza is so good" could be labeled as 'food', 'positive' or other attributes based on the created classification system. For such training data (ratings) to be reliable, researchers need to measure agreement between several raters. This is especially important when using crowd workers because the online environment includes risk factors such as bots pretending to be human raters, human raters engaged in fraudulent behavior, the difficulty of communicating tasks via the faceless online interface and many other challenges associated with crowdsourcing [3]. To calculate an inter-rater agreement score, each item (e.g., tweet) is classified by multiple raters into pre-existing categories.

Inter-rater agreement matters because it is used as a proxy of rating reliability [4], so when receiving a poor agreement, researchers and reviewers are left doubtful of the quality of the data. In contrast, when raters agree on items, it indicates that the data is trustworthy [5]. As noted by Mathet et al. [6], "when agreement is high, then the task is consistent and correctly defined, and the annotators can be expected to agree on another part of the corpus, or at another time, and their annotations therefore constitute a consensual reference". However, in practice, the agreement scores obtained by researchers often fell below the defined quality thresholds, as observed in computational linguistics [7]. Yet, meta-analyses focusing on social computing tasks are very scarce, and we were unable to locate any research particularly analyzing inter-rater agreement metrics in social computing research. To address this gap, we present the following research questions:

1) What are the inter-rater reliability metrics that social computing studies commonly use?
2) How well are these metrics performing considering the nature of social computing studies?

II. Literature Review

Crowdsourced labeling has developed in tandem with computational algorithms [8]. Although successful machine learning relies on the high quality of training data, data quality from the crowd remains challenging especially for tasks requiring more sophisticated human interpretation [9]. Therefore, evaluating the quality of crowdsourced ratings is a key concern in social computing research [3].

Because we want to include only social computing studies using crowd workers to label training data, we manually review the 196 results and choose 46 articles that dealt with using the crowd for labeling training data. These articles were analyzed in more detail, retrieving the following information from the articles: (1) number of rated items by the crowd

978-1-5386-9588-3/18/$31.00 ©2018 IEEE

TABLE I
INTER-RELIABILITY MEASURES IN COMPUTATIONAL SOCIAL SCIENCE

Measure	Average score	Frequency of studies (%)
Krippendorff's alpha	0.613	9 (30%)
Fleiss' kappa	0.600	19 (63%)
Cohen's Kappa	0.613	1 (3%)
Intraclass correlation coefficient (ICC)	N/A	1 (3%)

TABLE II
EXAMPLES OF 'MORE SUBJECTIVE' AND 'MORE OBJECTIVE' CROWD RATING TASKS FROM LITERATURE

More subjective	More objective
Level of abusive language [16]	Age, gender, ethnicity [17]
Ambiance of social media images [18]	Historical facts [19]
Suicide-related communication [20]	Topic of a news article [21]
Credibility of tweets [22]	Objects in Instagram images [23]
Debate performance [10]	Emergency information in tweets [24]

workers, (2) number of crowd raters per item, (3) number of categories used in the study, (4) description of categories (i.e., what was rated), (5) subjective or fact (i.e., the degree of subjectivity of the rating task), (6) reliability measure(s) used, and (7) score for the measure(s).

Table I shows the frequency of scores used and their averages across the studies. On average, the studies had 2,681 rated items with a high variation (sd = 5,073, cv = 1.89). Besides standard deviation, we also computed the coefficient of variation (cv) which is simply the standard deviation divided by the mean. This value is interpreted so that the closer it is to 1, the more variation the variable has. There were 4.06 raters per item on average with a low variation (sd = 1.22, cv = 0.30). Most commonly, there were 3 raters, which was the case in 45% of the studies reporting the number of raters.

Almost all articles (96%) report the number of rated items and the categories used. However, only 48% report how many raters per item were used. More importantly, 35% fail to report the inter-rater reliability metric used, and 48% do not report the value of the inter-rater agreement. In three papers, authors made the argument that kappa could not be calculated. For example, Diakopoulos and Shamma [10] noted that, since rating categories are not mutually exclusive, a rating reliability measure such as Fleiss' kappa is not appropriate.

In the ones reporting the agreement, the most common metric was Fleiss' kappa (63%). Krippendorff's alpha was used by 30% of the articles. Only one study, Kalaitzis et al. [11], reported using several inter-rater measures (Fleiss' kappa, ICC). The average reported inter-rater agreement across all metrics and studies was 0.60 (sd = 0.184). The coefficient of variation (cv = 0.305) does not indicate a high variation among the agreement scores.

Seven studies (15%) reported the inter-rater agreement by category. This practice, although neglected by the majority of the studies, is preferable as it reveals more information about the structure of agreement/disagreement [7]. For example, in one study the researchers were crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use, obtaining a high variation of rating scores (cv = 0.730), the lowest agreement being 0.057 (Fleiss' kappa) for 'Pleasure' category and the highest k = 0.943 for 'Written in English' category [12]. This example illustrates the importance of digging deeper into the source of the overall agreement score. Regarding the reported agreement scores, the studies tend to fall short of the suggested threshold values for the scores (average of all scores = 0.610). While a considerable number of works discusses the threshold of a good agreement [13], values below 0.5 are typically considered to indicate poor agreement. For example, Di Eugenio and Glass [14] suggest values of at least 0.67, while Artstein and Poesio [15] suggest target values ranging from 0.80 to 0.90. According to Bayerl and Paul [7], a common rule of thumb is to have at least 80% observed agreement. In this light, the agreement scores of social computing studies are not high.

III. EMPIRICAL INVESTIGATION WITH SOCIAL COMPUTING DATASETS

A. Overview

With this inquiry, we aim to clarify the degree of variation in agreement across several metrics defined by scholars, not only Fleiss' kappa and Krippendorff's alpha. This investigation may provide interesting information about the nature of the metrics in the context of using crowdsourced ratings for social computing tasks. The purpose of the mathematical notations is to illustrate the basic operating logic of the metrics. After this, we proceed to their application on the empirical datasets.

B. Observed Agreement

The simplest measure of agreement between raters is the *joint probability of agreement.*, also known as percentage agreement, observed agreement or simple agreement. If two raters agree in N_a out of total N items they rate, the observed agreement is given by :

$$p_o = \frac{\text{no. of items in which raters agree}}{\text{total no. of items}} = \frac{N_a}{N} \quad (1)$$

However, the measure fails to take into account the possibility that the agreement can take place by *chance.* To remove this effect of coincidental agreement, it is common to revert to some chance adjusted index, generically calculated as :

$$\text{Chance-adjusted Reliability} = \frac{p_o - p_c}{1 - p_c} \quad (2)$$

where the fraction of agreement due to chance (p_c) is removed from the agreed as well as total items. On one extreme, where there is no agreement due to chance (i.e. $p_c = 0$), Equation (2) reduces to (1); while if all agreement is due to chance, (i.e. $p_c = p_o$), the reliability drops to zero.

There exist several different approaches among researchers that deal with the question of "how many agreements take place by chance." In the following sub-sections some popular measures of chance-adjusted indices are discussed. Most of these indices use the generic form given by Equation (2), and only calculate the agreement differently.

C. Bennett et al.'s S-score

In 1954, Bennett et al. [25] proposed an S-score to represent the agreement between two raters where the effect of chance was nullified by subtracting the expected number of chance-agreements. From probability theory, if rater A rates an item into any one of the categories (say i of q), then there is a $1/q$ chance of rater B too rating that item as the same category (i),completely by chance. This uniform assumption gave rise to the chance-agreement (and subsequent S-score) as :

$$\begin{aligned} & p_c = 1/q \\ \implies \quad & S = \frac{p_o - 1/q}{1 - 1/q} \end{aligned} \tag{3}$$

D. Cohen's Kappa and Fleiss' Kappa

In 1960, Cohen [26] generalized the probability of raters assigning a certain category by chance. He presented the kappa coefficient of reliability, in which he proposed calculating p_c as :

$$\begin{aligned} & p_c = \frac{1}{N^2} \sum_{k=1}^{q} n_{k1}\, n_{k2} \\ \implies \quad & \kappa = \frac{p_o - \left(\sum_k n_{k1}\, n_{k2}\right)/N^2}{1 - \left(\sum_k n_{k1}\, n_{k2}\right)/N^2} \end{aligned} \tag{4}$$

where, of the total N items, rater 1 (or 2) rates n_{k1} (or n_{k2}) items as category k, for $k = 1, 2, \ldots, q$. This is one step sophistication of the S-score by assuming the raters predict different categories with unequal probabilities. The probabilities are calculated from their overall pattern of rating. Cohen assumed that the random category choices happen in the ratio of the categories chosen throughout all items.

While Cohen's kappa is used for 2 raters, it's extension for three or more raters was proposed by Fleiss [27] in 1971. With n raters assigning N items to one of the q categories, the basic formula for reliability comes from Equation (2) :

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{5}$$

where the fraction of items agreed is calculated as

$$\begin{aligned} & p_o = \frac{1}{N} \sum_{i=1}^{N} p_i \\ \text{with} \quad & p_i = \frac{1}{n(n-1)} \sum_{j=1}^{n} n_{ij}\,(n_{ij} - 1) \end{aligned} \tag{5a}$$

being the proportion of the number of pairs of raters in agreement about the i^{th} item, relative to all possible rater-pairs having rated i; and the fraction of items agreed by chance calculated as :

$$\begin{aligned} & p_c = \sum_{j=1}^{n} {c_j}^2 \\ \text{where} \quad & c_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \end{aligned} \tag{5b}$$

is the fraction of assignments given the j category by some or the other rater.

E. Scott's Pi

Soon after Bennnett et al., Scott [28] presented another chance-adjusted index. While the basic equation remains the same as eqn. (2), the chance-agreement is calculated in Scott's π as :

$$p_c = \left(\frac{m_1}{N}\right)^2 + \left(\frac{m_2}{N}\right)^2 \tag{6}$$

where $m_1 = \frac{1}{2}\,(n_{11} + n_{12})$ and $m_2 = \frac{1}{2}\,(n_{21} + n_{22})$ are the average number of items rated as category 1 and category 2 by a rater. This can be generalized to a multi-rater multi-category data using similar notions.

F. Krippendorff's Alpha

Krippendorff [29] assumed, very similar to Scott [28] that there is an inherent 'quota' assigned to each category, which the raters jointly fulfill. Krippendorff originally devised his coefficient to apply to multiple raters and multiple categories. However, for the simple 2-rater 2-category system, the chance agreement fraction can be calculated as :

$$p_c = \left(\frac{2m_1}{2N}\right)\left(\frac{2m_1 - 1}{2N - 1}\right) + \left(\frac{2m_2}{2N}\right)\left(\frac{2m_2 - 1}{2N - 1}\right) \tag{7}$$

When m_1, m_2 and N are sufficiently large, Equation (7) approaches (6), and α converges to π. For small number of items assigned to either category, Krippendorff's alpha works towards a more unbiased estimate. The more general case allows any number of raters, any number of categories, with weights assigned to category-pairs and raters free to skip items. For a system with N items to be rated into one of the q categories, out of which N' were rated by more than one rater, the basic formula for reliability is the same as any chance-adjusted index of Equation (2).

$$\alpha = \frac{p_o - p_c}{1 - p_c} \tag{8}$$

where the fraction of items agreed (p_o) is calculated as

$$\begin{aligned} & p_o = p'_o\,(1 - \epsilon_n) \;+\; \epsilon_n \\ \text{with} \quad & p'_o = \frac{1}{N'} \sum_{i=1}^{N'} \sum_{k=1}^{q} \frac{r_{ik}\,(r^*_{ik} - 1)}{r_i\,(r^*_i - 1)}, \\ \text{and} \quad & \epsilon_n = \frac{1}{N'\,\bar{r}}; \end{aligned} \tag{8a}$$

r_{ik} being the number of raters who rated item i as category k,

$r^*_{ik} = \sum_l w_{kl}\, r_{il}$ the weighted sum of raters who rated item i as category k,
$r_i = \sum_k r_{ik}$ the number of ratings recorded for item i (excluding skips), and
$\bar{r} = \frac{1}{N'} \sum_i r_i$ being the average number of ratings given on an item.
Similarly, the fraction of items agreed by chance is:

$$p_c = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl}\, \pi_k\, \pi_l \tag{8b}$$

$$\text{with} \qquad \pi_k = \frac{1}{N'} \sum_{i=1}^{N'} \frac{r_{ik}}{\bar{r}}$$

denoting the probability of finding a subject and a rater who classifies the subject into category k.

G. Gwet's Gamma

Gwet [30] sought a further level of sophistication to calculate the chance-agreement, using a hybrid approach based on category as well as average distribution. It is claimed to be resistant to the following two paradoxes characteristic of kappa indices [31]:

- the formulation of chance-agreement in a kappa index tends to convert a high observed agreement fraction to a relatively low value of κ; and
- imbalanced class problems tend to inflate kappa unusually.

Commonly denoted by $AC1$, this coefficient is originally devised for multi-rater and multi-category cases. However, for the sake of comparison, the two-rater two-category chance-agreement may be derived as:

$$p_c = \left(\frac{m_1}{N}\right)\left(\frac{m_2}{N}\right) + \left(\frac{m_2}{N}\right)\left(\frac{m_1}{N}\right) \tag{9}$$

This formulation of inter-rater reliability is adjusted both for chance and also for class imbalance — a predominant case in many real-world data, including crowdsourced ratings. For multiple raters, multiple categories and missing data, Gwet's $AC1$ index calculates the fraction of chance agreement as the tendency of raters to agree on hard-to-rate items. It is calculated as the product of the probabilities of agreeing when the rating is random, and of finding a hard-to-rate subject:

$$p_c = \frac{1}{q(q-1)} \sum_{k=1}^{q} \pi_k (1 - \pi_k) \tag{10}$$

with the expression for π_k given in Equation (8b).

H. Empirical Datasets

We then apply the inter-rater measures to four datasets collected via crowdsourcing. These datasets represent typical labeling tasks in social computing: text and image classification, with varying degree of subjectivity. The datasets were obtained from researchers who have published social computing researcher and were willing to share their datasets and represent a typical social computing tasks. The datasets are summarized in Table III.

In Task 1, crowd workers were asked to label /r/AskReddit comments as either toxic or not toxic. In this context, a comment is considered toxic if it is "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." If a crowd worker labeled a comment toxic, they were asked to rate how toxic the comment is: slightly toxic (rated as 1) or highly toxic (rated as 2).

In Task 2, the purpose was to determine the age, gender, and ethnicity of Twitter users. The crowd workers were asked to provide this information after reviewing the profile picture provided. If the picture contained a human face, crowd workers were asked to choose "Yes"; otherwise, they should choose "No". When selecting "Yes", they were asked to evaluate gender, ethnicity, and age of the person. The workers were asked to choose 'No' when the picture included non-human faces, drawings or illustrations.

In Task 3, workers were asked to identify images for humanitarian aid. The purpose was to understand the type of information shared in an image. The data was collected from Twitter during the Hurricane Irma in 2017. The workers were instructed to categorize images into one of the following categories: Infrastructure and utility damage; Vehicle damage; Rescue, volunteering, or donation effort; Injured or dead people; Affected individuals; Missing or found people; Other relevant information; and Not relevant or cannot judge.

In Task 4, crowd workers were asked to evaluate advertising copywriting sentiments. Four persuasive techniques were applied to short text ads, and the task was to tell how the text makes the worker feel. The techniques were Emotional ("the text makes you feel emotional"), Social pressure ("the ad makes you feel social pressure to do something"), Fear of missing out ("the ad makes you fear missing out on a great opportunity"), and Neutral "the ad has gives you no special feeling"). The purpose of this study was to analyze the use of persuasion tactics in different channels, including SMS, mobile app, Facebook, and email.

Overall, the tasks represent different types of classification tasks in social computing. We next proceed to calculate the reliability scores using the functions provided by the MATLIB package mReliability [32], available in GitHub[1]. The equations used by this software package correspond to the ones presented in Section III. We calculate the agreement scores both by class and by dataset. Table IV shows the reliability scores obtained for the four datasets.

Observed agreement is the highest across all the datasets. This is to be expected, given that chance-adjustment categorically reduces the agreement score. It is worth noting that while the observed agreement is high, Fleiss's Kappa is rather low; this problem was heavily discussed by Feinstein and Cicchetti [33], associating the low Kappa score with class imbalance. For example, if 80% of all the items were assigned the 'non-toxic' category, then $4/5^{\text{th}}$ of the random choices should also be 'non-toxic'. This pushes the chance-agreement fraction to 0.68 in Cohen's kappa, contrasted to 0.5

[1] https://github.com/jmgirard/mReliability

TABLE III
FOUR SOCIAL COMPUTING DATASETS

	DATA 1	DATA 2	DATA 3	DATA 4
Task name	"Is This Askreddit Comment Toxic? If So, How Toxic Is It?"	"Categorizing Faces From Twitter Profile Pictures"	"Identify Images Useful For Humanitarian Aid"	"Evaluate Ad Copy Texts"
Target to classify	Toxicity of a comment	Age, gender and ethnicity of a user	Crisis-related content in images	Sentiment of advertising texts
No. of ratings	30,008	3,000	8,748	200
No. of raters	3	3	3	50

TABLE IV
INTER-RATER RELIABILITY SCORES FOR THE DATASETS

	DATA 1	DATA 2	DATA 3	DATA 4
Obs. Agreement	0.71	0.75	0.684	0.292
Gwet's Gamma	0.48	0.70	0.646	0.062
Bennett's S Score	0.42	0.69	0.638	0.056
Scott's Pi	0.30	0.64	0.578	0.040
Krippendorff's Alpha	0.31	0.64	0.578	0.045
Fleiss' Kappa	-0.01	0.64	0.578	0.040

in Brennett et al.'s assumption of choosing a random category with equal probability. On the other hand, Gwet's gamma and other average-distribution approaches handle such class imbalance differently and potentially out-perform the Kappa statistic measurements.

We can see from Figure 1 that, in the case of our data, raters could most easily agree about gender and if a social media comment is toxic (a large share of 3 raters agreeing, blue bars in Figure 1). In contrast, age and intensity of toxicity of a comment were the hardest to agree upon (a large share of no agreements, grey bars in Figure 1). Looking at the labels across all categories, we find that the annotators found it easiest to agree on the ethnicity of white profile pictures (27% of total observations were where three raters agreed the person was White), while the hardest was to agree on Asian (3%, compared to Black: 12%).

Figure 2 illustrates the variation among agreement scores across all the classes. As can be seen, the agreement scores tend to co-vary. The major exception is Fleiss' kappa which penalizes strongly the crisis-related content and the perceived intensity of hate. Overall, the degree of toxicity and ad sentiment are hardest to rate. This can be explained by the inherent subjectivity of these tasks: there is a universal truth of how an ad makes a person feel, or how hostile one perceives a comment. Moreover, age and ethnicity seem to be more difficult to rate than gender, which suggests that as subjectivity increases, the harder it is to reach an agreement. Although age has an objectively correct value, people of the same age can look very different, resulting in varying *perceptions of age*. The degree of toxicity is similar: people generally agree whether a comment is hateful or not, but when asked how hateful, there is large disagreement, arising from perceptual differences of a fine-grained classification task.

Looking at the correlation between scores, we find that Gwet's gamma has the highest correlation with the observed agreement (r = 0.95), which means this score penalizes the least from the baseline of the simple agreement. In turn, Fleiss' kappa (r = 0.54) has the lowest correlation with the observed agreement. Looking at correlations between the tasks, we find that the scores between Tasks 2 and 3 have the highest pairwise correlation (r = 0.992). Both of these tasks were image classification tasks. The lowest correlation coefficient is between Task 1 (toxic comments) and 4 (ad sentiment) (r = -0.221). To understand the variation of the metrics better, we next perform a sensitivity analysis by varying the number of items, raters and categories.

IV. SENSITIVITY ANALYSIS

We analyze the variation and robustness of the different metrics to parametric deviations commonly encountered in the ratings obtained from the crowd. Sensitivity analysis may be defined as the study of how uncertainty in a measured quantity can be attributed to different sources of uncertainty in the parameters on which the quantity depends [34]. This section analyses and studies the sensitivity of the different metrics to parameters affecting agreement scores. In particular, we wish to investigate the effects of increasing the number of (a) rated items, (b) raters, and (c) categories. The analysis essentially consists of the following experiments performed with an objective to test the variation of reliability measures with the above factors.

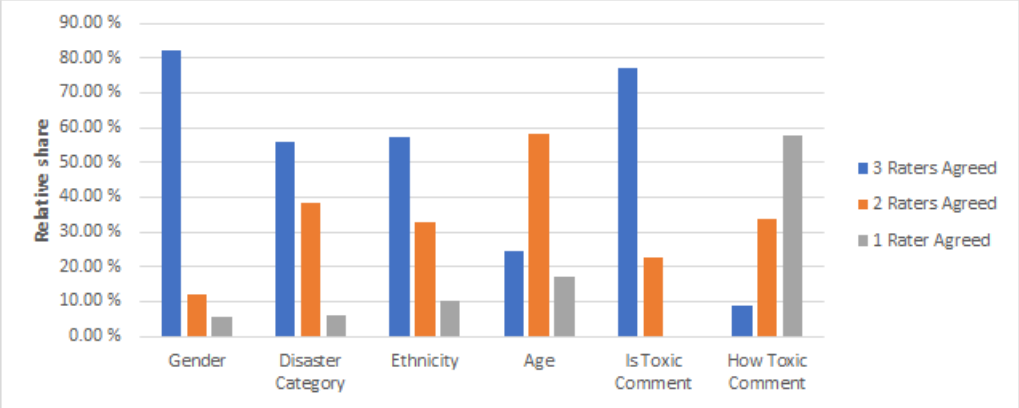


Fig. 1. Share of Times Raters Agreed on an Item in a Given Category

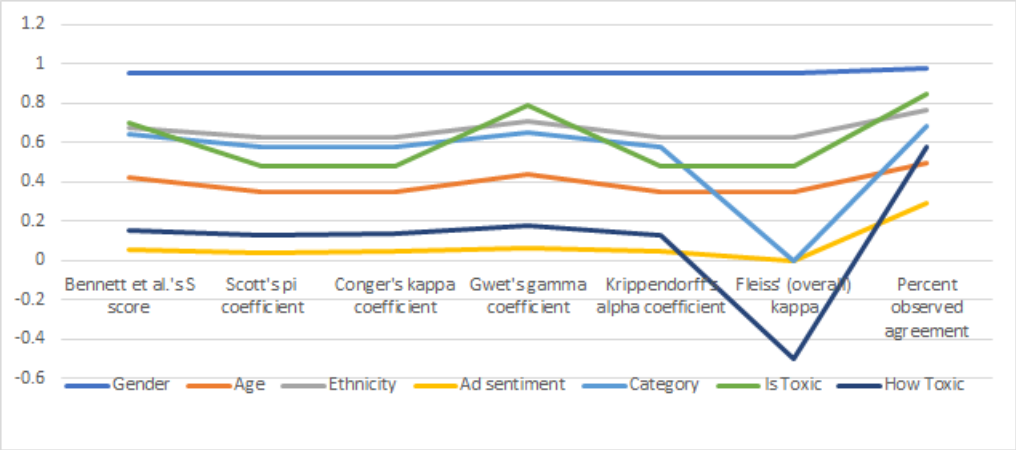


Fig. 2. Variation of Metrics Across Classes

1) **Raters :** DATA 3 had a total of 2,236 items, of which 40 items were rated by at least 30 raters. 4 sub-samples were prepared with 40 items each, rated by 3, 5, 10 and 30 raters.
2) **Items :** From DATA 1 with 10,000 items, three subsamples were taken with 500, 1,500, 5,000 items each. These three samples, along with the full (10,000 items) dataset contributed to the test setup for the items experiment.
3) **Categories :** DATA2 originally has 8 categories for 'Age'. These 8 categories were hierarchically merged to form 4 broader categories with balanced number of items in each category. The two datasets with 4 and 8 categories served as the other experimental data for the categories experiment.

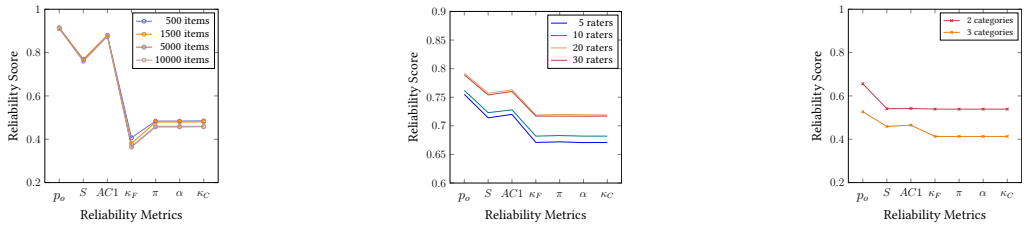
The simple percent agreement and the chance-adjusted agreement scores listed in Section III were calculated for these 10 different datasets with varying items, raters, and categories. The samples were drawn from the main dataset using pseudo-random numbers generated by a Fortran 90 script. The corresponding variation of scores is given in Figure 3.

We find that most reliability scores for a particular task are quite consistent across the number of items, provided there are at least a few hundred items to be rated (Figure 3 (a)). From Figure 3, it is evident that the reliability scores are most sensitive to categories, somewhat less sensitive to the number of raters, and the least sensitive to the number of items. Naturally, this observation applies to the parametric variation within our data. For example, the present study is unable to shed any light on what would happen to Scott's pi should the ratings were done by 100 raters instead of 2. However, the variables were selected in a range that is most relevant for crowd ratings.

We also calculate the sensitivity index which is the slope of the line that best fits the plot of metrics scores versus the value of the parameters taken on a $\log_2$ scale :

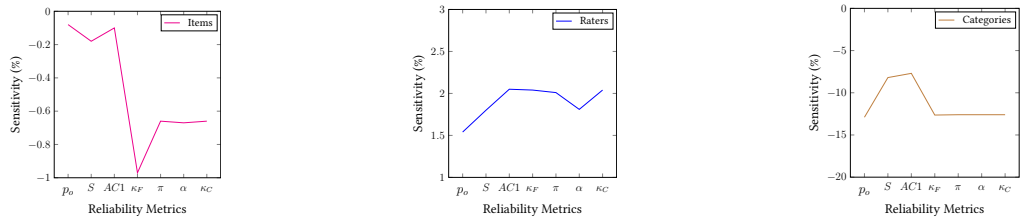
$$d = \frac{\sum (x_i - \overline{x})\,(y_i - \overline{y})}{\sum (x_i - \overline{x})} \tag{11}$$

where the log of the number of items/raters/categories (with base 2) are considered on the x-axis, and the inter-rater-reliability scores are considered on the y-axis. The sensitivity index d has been presented as a percent in this section. The slope depicts the rate at which reliability increases when the number of items, raters or categories is doubled. For example, Figure 4 (a) and (b) can be interpreted so that Gwet's $AC1$



(a) (b) (c)

Fig. 3. Variation of the Different Reliability Metrics With Varying (a) Items, (b) Raters, and (c) Categories



(a) (b) (c)

Fig. 4. Sensitivity of the different reliability metrics with (a) items, (b) raters, and (c) categories

decreases by 0.1% (or increases by 2%) if the number of items (or raters) was doubled in case of the performed experiments.

From Figure 4 it is evident that while the metrics have a slight positive sensitivity to a number of raters, a variation of the metrics with a number of items is negative and quite low. Sensitivity to the number of categories is comparatively much higher and on the negative side.

In particular Brennett's S-score and Gwet's AC1 coefficient shows comparatively less sensitivity to the number of categories. However, this trend was observed in a limited scope of experiments, and more experiments should be done before a general conclusion can be reached.

V. Conclusion and Discussion

To summarize, our research provides five main findings:

- Nature of rating tasks in social computing tend to be more subjective than objective.
- Most social computing studies do not adequately report the information needed to evaluate the agreement of crowd workers.
- The reported agreement scores in social computing studies are not high, averaging at around 0.60 for both Kappa and Alpha metrics.
- The more subjective the task, the worse the agreement tends to be, regardless of the used metric.
- Different metrics penalize differently compared to observed agreement - Fleiss' kappa seems to penalize relatively much, while Gwet's gamma penalizes relatively little. Therefore, using Gwet's gamma, or even the simple percentage agreement might be appropriate.

Three key points emerge for discussion. First, even if the inter-rater score is low, the crowd raters may still be reliable, i.e., giving honest responses. A practical challenge then is how to identify between the two, determining when the disagreement is true due to bad guidance or rater quality and when due to the subjectivity of the task.

Second, the commonly used metrics, namely Fleiss' kappa and Krippendorff's alpha, use chance-adjustment. However, there is no "chance" involved when the rater is giving his or her true opinion on a subjective matter. In fact, using chance-adjusted reliability scores is misleading in this case as they incorrectly bias away from true values. Coincidentally, this leads us to venture further away from perceiving crowd as an anonymous mass of people, as it was originally intended [35], into perceiving crowd workers more as *opinionated individuals*. Such a step could be useful for computational social scientists, even though it requires developing new sampling strategies and finding out more about the crowd raters.

Finally, although most researchers did report the basic numbers, from the reviewed papers, we conclude that crowd labeling is not always properly reported by researchers. This hinders replication and evaluation of the use of crowd annotators. Specifically, only a few studies reported category-specific agreements despite encouragements in the prior literature to do so [7]. Future research utilizing crowd raters should (a) report the sources of disagreement, and (b) set the goal for agreement score according to the level of task subjectivity.

Acknowledgment

The authors thank Dr. Muhammad Imran from Qatar Computing Research Institute for giving access to datasets.

References

[1] F. R. A. Neto and C. A. Santos, "Understanding crowdsourcing projects: A systematic review of tendencies, workflow, and quality management," *Information Processing and Management*, vol. 54, no. 4, pp. 490 – 506, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457317308476

[2] D. Reidsma and J. Carletta, "Reliability measurement without limits," vol. 34, no. 3, pp. 319–326, 2008-09. [Online]. Available: http://dx.doi.org/10.1162/coli.2008.34.3.319

[3] O. Alonso, C. C. Marshall, and M. Najork, "Debugging a crowdsourced task with low inter-rater agreement," in *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '15. ACM, 2015, pp. 101–110. [Online]. Available: http://doi.acm.org/10.1145/2756406.2757741

[4] K. Krippendorff, "Reliability," 2017-04-24. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118901731.iecrm0210

[5] M. Joyce. (2013) Picking the best intercoder reliability statistic for your digital activism content analysis. [Online]. Available: http://digital-activism.org/2013/05/picking-the-best-intercoder-reliability-statistic-for-your-digital-activism-content-analysis/

[6] "The unified and holistic method gamma for inter-annotator agreement measure and alignment," vol. 41. [Online]. Available: http://dx.doi.org/10.1162/COLI00227

[7] "What determines inter-coder agreement in manual annotations? a meta-analytic investigation," vol. 37. [Online]. Available: http://dx.doi.org/10.1162/COLIa00074

[8] "Using semantic similarity to reduce wrong labels in distant supervision for relation extraction."

[9] F.-Y. Hsu, H.-M. Lee, T.-H. Chang, and Y.-T. Sung, "Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques," *Information Processing and Management*, vol. 54, no. 6, pp. 969 – 984, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457317308245

[10] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. ACM, 2010, pp. 1195–1198. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753504

[11] A. Kalaitzis, M. I. Gorinova, Y. Lewenberg, Y. Bachrach, M. Fagan, D. Carignan, and N. Gautam, "Predicting gaming related properties from twitter profiles," in *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, 2016, pp. 28–35.

[12] N. Alvaro, M. Conway, S. Doan, C. Lofi, J. Overington, and N. Collier, "Crowdsourcing twitter annotations to identify first-hand experiences of prescription drug use," vol. 58, pp. 280–287, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046415002415

[13] A. Tordai, J. van Ossenbruggen, G. Schreiber, and B. Wielinga, "Let's agree to disagree: On the evaluation of vocabulary alignment," in *Proceedings of the Sixth International Conference on Knowledge Capture*, ser. K-CAP '11. ACM, 2011, pp. 65–72. [Online]. Available: http://doi.acm.org/10.1145/1999676.1999689

[14] B. D. Di Eugenio and M. Glass, "The kappa statistic: A second look," vol. 30, no. 1, pp. 95–101, 2004. [Online]. Available: https://doi.org/10.1162/089120104773633402

[15] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," vol. 34, no. 4, 2008. [Online]. Available: http://dx.doi.org/10.1162/coli.07-034-R2

[16] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16. International World Wide Web Conferences Steering Committee, 2016, pp. 145–153.

[17] X. Chen, Y. Wang, E. Agichtein, and F. Wang, "A comparative study of demographic attribute inference in twitter." in *Ninth International AAAI Conference on Web and Social Media*, vol. 15, 2015, pp. 590–593.

[18] D. Santani, R. Hu, and D. Gatica-Perez, "InnerView: Learning place ambiance from social media images," in *Proceedings of the 2016 ACM on Multimedia Conference*, ser. MM '16. ACM, 2016, pp. 451–455. [Online]. Available: http://doi.acm.org/10.1145/2964284.2967261

[19] N.-C. Wang, "Crowdnection: Connecting high-level concepts with historical documents via crowdsourcing," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '16. ACM, 2016, pp. 146–151. [Online]. Available: http://doi.acm.org/10.1145/2851581.2890377

[20] P. Burnap, W. Colombo, and J. Scourfield, "Machine classification and analysis of suicide-related communication on twitter," in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, ser. HT '15. ACM, 2015, pp. 75–84. [Online]. Available: http://doi.acm.org/10.1145/2700171.2791023

[21] C. Orellana-Rodriguez, D. Greene, and M. T. Keane, "Spreading ones tweets: How can journalists gain attention for their tweeted news?" vol. 3, no. 1, pp. 16–31, 2017. [Online]. Available: http://www.nowpublishers.com/article/Details/JWS-0009

[22] K. R. Canini, B. Suh, and P. L. Pirolli, "Finding credible information sources in social networks based on content and social structure," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 2011, pp. 1–8.

[23] Y. Hu, L. Manikonda, and S. Kambhampati, "What we instagram: A first analysis of instagram photo content and user types," 2014.

[24] V. Hester, A. Shaw, and L. Biewald, "Scalable crisis relief: Crowdsourced SMS translation and categorization with mission 4636," in *Proceedings of the First ACM Symposium on Computing for Development*, ser. ACM DEV '10. ACM, 2010, pp. 15:1–15:7. [Online]. Available: http://doi.acm.org/10.1145/1926180.1926199

[25] E. M. Bennett, R. Alpert, and A. C. Goldstein, "Communications through limited-response questioning," *Public Opinion Quarterly*, vol. 18, no. 3, pp. 303–308, 1954. [Online]. Available: https://doi.org/10.1086/266520

[26] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. [Online]. Available: https://doi.org/10.1177/001316446002000104

[27] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[28] W. A. Scott, "Reliability of content analysis:the case of nominal scale coding," *Public Opinion Quarterly*, vol. 19, no. 3, pp. 321–325, 1955. [Online]. Available: https://doi.org/10.1086/266577

[29] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970. [Online]. Available: https://doi.org/10.1177/001316447003000105

[30] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *British Journal of Mathematical and Statistical Psychology*, vol. 61, no. 1, pp. 29–48, 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1348/000711006X126600

[31] A. R. Feinstein and D. V. Cicchetti, "High agreement but low kappa: I. The problems of two paradoxes," *Journal of Clinical Epidemiology*, vol. 43, no. 6, pp. 543 – 549, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/089543569090158L

[32] J. Girard, "mReliability: Reliability analysis in MATLAB," 2016. [Online]. Available: http://mreliability.jmgirard.com

[33] A. R. Feinstein and D. V. Cicchetti, "High agreement but low kappa: I. the problems of two paradoxes," vol. 43, no. 6, pp. 543–549, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/089543569090158L

[34] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Experimental Designs*. Wiley-Blackwell, 2008, ch. 2, pp. 53–107.

[35] J. Howe, "The rise of crowdsourcing," vol. 14, no. 6, pp. 1–4, 2006.