Online Hate Ratings Vary by Extremes: A Statistical Analysis

Joni Salminen
Qatar Computing Research Institute, HBKU;
and Turku School of Economics
Doha, Qatar
jsalminen@hbku.edu.qa

Hind Almerekhi
Hamad Bin Khalifa University
Doha, Qatar
hialmerekhi@hbku.edu.qa

Ahmed Mohamed Kamel
Cairo University
Cairo, Egypt
ahmedm.kamel@pharma.cu.edu.eg

Soon-gyo Jung
Qatar Computing Research Institute,
HBKU
sjung@hbku.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute, HBKU
Doha Qatar
bjansen@hbku.edu.qa

ABSTRACT

Analyzing 5,665 crowd ratings on 1,133 social media comments, we find that individuals tend to agree on the extremes of a hate rating scale more than in the middle when evaluating the hatefulness of online comments. The agreement is higher for less hateful comments and lowest on moderately hateful comments. The results have implications for researchers developing machine learning models for online hate processing, as the extreme classes are likely to require fewer annotations for reaching statistical stability. Our findings suggest that the models developed in this domain should consider the distributions of hate ratings rather than average hate scores.

CCS CONCEPTS

• Applied computing → Law, social and behavioral sciences → Sociology

KEYWORDS: Online hate, toxicity, ratings; interpretation; crowdsourcing

ACM Reference format:
Joni Salminen, Hind Almerekhi, Ahmed Mohamed Kamel, Soon-Gyo Jung, Bernard J. Jansen. 2019. Online Hate Ratings Vary by Extremes: A Statistical Analysis. In *Proceedings of ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR'19) March 10–14, 2019, Glasgow, United Kingdom.* ACM, New York, NY, USA, 5 pages. DOI: https://doi.org/10.1145/3295750.3298954

1 INTRODUCTION

Online hate research is important because hate has become commonplace in many social networks, communities, and websites [11, 13, 16]. Overall, online hate has a detrimental effect on the health of online communities [12], degrading the user experience of all users and specifically vulnerable groups [10].

For these reasons, researchers are working continuously to improve automatic methods for online hate processing, including detection, classification, manipulation, and moderation [16]. A major part of this work is the creation of training data that will be used to train the algorithms to better detect and score online comments for the hatefulness of their content.

However, interpretation of hatefulness varies among individuals [19], hindering the use of approaches that assign one hate score to an online comment. Thus, the distribution of online hate ratings is important for the creation of better models for online hate processing.

In this research, we pursue two research questions that relate to the nature of online hate ratings:

- **RQ1:** Are crowd raters agreeing more on comments that are considered as either very hateful or very little hateful, compared to moderately hateful comments?
- **RQ2:** Are crowd raters more likely to agree when a comment is more hateful or less hateful on average?

To answer these questions, we collect data from a crowdsourcing platform and conduct a statistical analysis.

2 RELATED LITERATURE

The high prevalence of online hate has urged research in detecting, classifying, and scoring hateful comments [16]. Typically, these efforts involve machine learning trained with crowdsourced ratings [21]. An example is Perspective API, a tool by Alphabet for scoring toxicity of online comments, that is trained using crowd annotators [9]. In previous research, toxicity has been seen conceptually equivalent to online hate [16].

However, the major issue in the existing approaches is that they consider hate as an aggregate value (e.g., a comment is “0.85 hateful”) rather than the empirical distribution of the ratings. A notable exception is Wulczyn et al. [22] that utilizes empirical distribution, however, only using binary ratios (e.g., the comment is “[0.8] toxic and [0.2] non-toxic”). Previous research has shown many sources for the variation of interpretation for online hate. For example, the hate interpretation may depend on the context [2], such as community norms and specific use of language [15], so

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298954

that the same language is perceived hateful in one context and non-hateful in another context.

Moreover, the interpretation of hatefulness varies by the individual raters, and there is also evidence suggesting that people from different countries perceive hatefulness of the same online comments differently [19].

3 METHOD

3.1 Data collection

Similar to Chatzakou et al. [3], we use a dictionary-based method to retrieve hateful comments. To build a dictionary, we compiled the list of hate words by using a two-staged process: (1) first, we searched for lists of profanity and/or swear words available online, finding four lists that appeared valid for our purpose[1].

Second, (2) we expanded these lists by identifying more hateful words using 500 randomly sampled comments from the online comments dataset collected by Salminen et al. [18], which originates from a YouTube channel of a major online news and media company. We added the identified hateful words and phrases in our hate keyword dictionary.

Searching with the dictionary, we retrieved a random sample of 1,133 comments from the said dataset that contain hateful language. With these comments, we proceed to crowdsourcing. We set up a data collection task on CrowdFlower (currently called Figure Eight), asking the crowd workers to rate the hatefulness in each comment, with the options Very hateful (4), Moderately hateful (3), Slightly hateful (2), or Not hateful at all (1). Hatefulness was defined as follows: "*Hateful comment is something that aims to hurt another person or contains bad language.*"

A total of 5,665 ratings were collected, so that each of the 1,133 YouTube comments was rated by 5 reviewers. We undertook several measures to ensure the quality of work inputs:

- **Max Judgments Per Rater:** *50 judgments*, to limit the impact of individual raters' bias on the results.
- **Minimum Time:** *25 seconds*; a page has 5 comments and we expect it would take a rater at least 5 seconds to read and rate each.
- **Quality Level:** *Level 2 ("Higher Quality")*, consisting of a smaller group of more experienced, higher accuracy raters on CrowdFlower.
- **Judgments Per Row:** *5 judgments*, to get more information on the preferences, as is suggested for more subjective crowd classification tasks [1, 17].
- **Price per Judgment:** *4 cents USD*, a 33% increase in the price suggested by CrowdFlower.

3.2 Interrater agreement

The interrater agreement was first calculated using Krippendorff's alpha (K alpha) that is a good measure of agreement when an ordinal scale is used for rating. K alpha can take a value that from 0 to 1, with 0 indicating perfect disagreement and 1 perfect agreement. However, the values of K alpha tend to show less agreement than percentage agreement (PA), since K alpha corrects for the expected agreement by chance. This is problematic when the expected agreement due to chance is high, as this may bias the calculation of K alpha [8]. This phenomenon is known as Krippendorff's alpha paradox [6].

Thus, an additional measure was used, namely, Gwet's AC2, that can be used in the presence of high expected agreement by chance, as it does not assume independence between raters [7]. Like K alpha, Gwet's AC2 can range from 0 to 1, with higher values indicating higher agreement between the raters.

3.3 Analytical approach

RQ1: The average hatefulness score (AHS) for each comment was calculated by averaging the ratings for the five raters. A threshold was then identified and used to separate the comments into lower and upper extremes and moderate hatefulness groups.

The data was divided into three groups based on the 25th percentile and the 75th percentile. Comments with a score below the 25th percentile or above 75th percentiles were identified as extremes. K alpha and PA were calculated for each of the three groups. Scores were compared between the three groups as well as to the overall K alpha of the whole data.

Bootstrapping (with 1,000 resamples) was used to obtain the 95% confidence interval (Bias corrected and accelerated) for K alpha and PA for each of the three groups, as and well as for the overall score. For Gwet's AC2, bootstrapping was not needed, as the metric incorporates the standard error for the coefficient that can be used to calculate the 95% confidence interval.

RQ2: We used an approach similar to RQ1 to compare the agreement for more hateful comments to the agreement for less hateful comments. The toxicity score that captures the degree of online hate in a comment was calculated for each comment. The median toxicity score (50% percentile) was used to classify the data into more hateful (upper 50%) and less hateful (lower 50%) comments. Agreement was calculated for each group (more hateful vs. less hateful) and scores were compared between the groups. Bootstrapping was used to obtain the 95% confidence interval for K alpha and PA.

4 RESULTS

4.1 Descriptive statistics

Table 1 shows the agreement measures for all rated comments. The overall agreement among participants is only 22.6%. The overall K alpha is 0.55, which is considered moderate. Gwet's AC2 is 0.369, which is also considered moderate.

Table 1: Measures of interrater agreement across all comments (n = 1133)

Measure	Results [95% CI]
Krippendorff's alpha	0.55 [0.48 – 0.62]
Percentage agreement (%)	22.6 [20.21 – 25.15]
Gwet's AC2	0.369 [0.344 -0.394]
Toxicity score (median [IQR])	2 [1.2 – 3]

[1] http://www.bannedwordlist.com/lists/swearWords.txt; https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en; https://www.frontgatemedia.com/a-list-of-723-bad-words-to-blacklist-and-how-to-use-facebooks-moderation-tool/; http://onlineslangdictionary.com/lists/most-vulgar-words/

4.2 RQ1: Extreme vs. medium perceptions

The median toxicity score is 2, which means that 50% of the comments have an AHS lower than 2, while the remaining 50% have an average hatefulness score greater than 2. The interquartile range is 1.2 – 3, implying that 25% of the comments have an AHS below 1.2, while 25% have an AHS greater than 3.

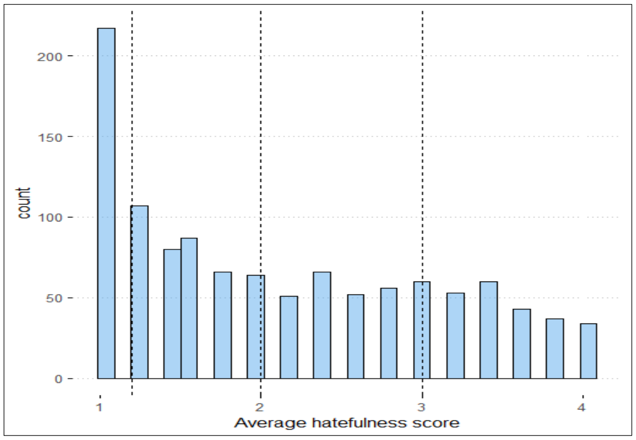

Figure 1: Distribution of average hatefulness score

From Figure 1, we can see that the AHS is right-skewed. Thus, interquartile range (IQR) is a good option to separate the comments into three groups, based on the 25th and 75th percentiles. Values below the 25th percentile or above the 75th percentile are identified as lower and upper extremes, respectively. Table 2 compares agreement measures across moderately and extremely hateful comments.

Table 2: Agreement across groups (95% CI)

Group	K alpha	PA	AC2
Lower extreme (N = 324)	0 [0 – 0.46]	66 [61.42 – 71.6]	0.85 [0.82 – 0.88]
Moderate (N = 582)	0.08 [0 – 0.18]	0.86 [0.17 – 1.72]	0.25 [0.22 – 0.29]
Upper extreme (N = 227)	0.02 [0 – 0.15]	14.98 [10.57 – 19.38]	0.76 [0.73 – 0.79]

K alpha and PA were calculated for each of the three groups. Bootstrapping (using 1000 resamples as recommended by Efron [5]) was used to calculate the 95% confidence interval.

Results for K alpha show some conflicts with the percentage agreement measure. The K alpha for the extreme two groups was very low compared to PA. This can be explained by examining the distribution of ratings for the three groups (see Table 3).

Table 3 shows that most raters in Group 1 rated the comment as either 1 (n = 1513, 93.4%) or 2 (n = 107, 6.6%). The same was true for Group 3 where most raters rated the comments as either 3 (n = 347, 30.57%) or 4 (n = 712, 62.73%). This makes K alpha a less reliable indicator in such a case, since it compares the observed agreement to the expected agreement by chance (which will be very high in these two cases). This will result in a lower K alpha than expected. Thus, PA and AC2 are more appropriate measures to answer the first research question.

Table 3: Distribution of reviews across the three groups

Group	1	2	3	4
Lower extreme	1513 (93.4%)	107 (6.6%)	0 (0%)	0 (0%)
Moderate	1082 (37.18%)	691 (23.75%)	347 (30.57%)	328 (11.27%)
Upper extreme	16 (1.41%)	60 (5.29%)	347 (30.57%)	712 (62.73%)

Results in Table 2 show that PA was highest for the lower extreme group (PA = 66%, 95% CI = 61.42 – 71.6). This means that PA in the lower extreme is 66%. The 95% confidence interval 61.42% - 71.6% means that 95% of the intervals constructed will contain the true agreement score.

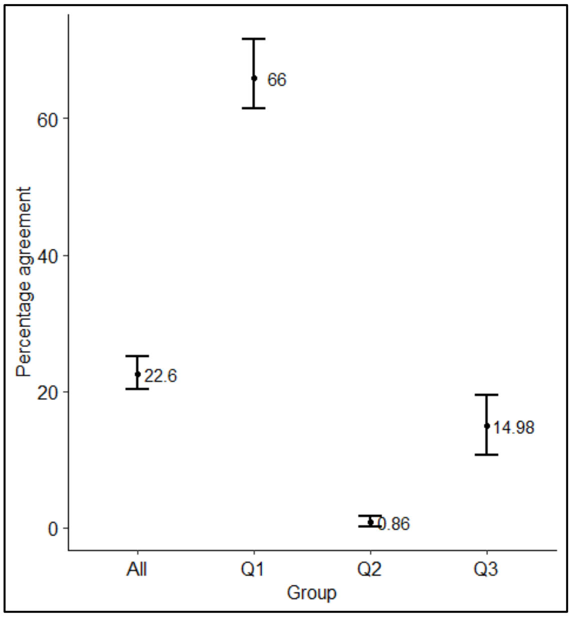

Figure 2: Percentage agreement across the whole data and per group

The PA in the upper extreme was 14.98% (95% CI = 10.57 – 19.38) which is relatively low compared to the lower extreme but high when compared to the comments with moderate hatefulness (Group 2) where the agreement was nearly zero (PA = 0.86, 95% CI = 0.17 – 1.72). Results from AC2 were in line with PA, showing the highest agreement for the lower extreme. The agreement is lower in the upper extreme than in the lower extreme, while the moderate hatefulness group showed the lowest AC2.

These findings show that raters agree most when they review comments that are in the *lower extreme*. They agree to a lower extent when the comment is in the *upper extreme*, and they rarely agree when the comment has *moderate hatefulness*.

Results from Figure 2 show that confidence intervals do not overlap, which indicates that the differences in PA between groups are statistically significant. Group 1 (least toxic comments) had the highest PA. Group 3 (most toxic comments) had the second highest, while moderately toxic comments had the lowest PA.

These results show that extreme hatefulness have more agreement than medium hatefulness perceptions.

4.3 RQ2: More vs. less hateful comments

We first identify the threshold for more and less hateful comments. The median toxicity score was 2 (Table 1), which means that comments with an AHS below 2 were considered as less hateful while comments with an AHS above 2 were considered as more hateful comments.

Using these thresholds, we compare agreement across more and less hateful comments

Table 4: Agreement across more and less hateful comments (95% CI)

Group	K alpha	PA	Gwet's AC2
More hateful (N =621)	0.17 [0.07 – 0.27]	7.62 [5.27 – 9.96]	0.88 [0.79 – 0.83]
Less hateful (N =512)	0.10 [0 – 0.23]	34.9 [30.76 – 39.37]	0.41 [0.37 – 0.46]

Agreement was calculated for each of the two groups. Bootstrapping (with 1000 resamples) was used to calculate the 95% confidence interval for K alpha and PA. The same issue that we identified with K alpha in RQ1 was identified for less hateful comments, since a few comments were rated as either as 3 or 4, while the majority were rated as 1 or 2. This increases the expected agreement and lowers the K alpha of less hateful compared to the more hateful group. The distribution of responses, however, shows that the agreement is much higher among less hateful comments, since only a small number of responses were 3 and 4, while the majority were 1 and 2 (75.62% and 14.43%, respectively).

Table 5: Distribution of reviews across the two groups

Group	1	2	3	4
Less hateful	2348 (75.62%)	448 (14.43%)	263 (8.47%)	46 (1.48%)
More hateful	263 (10.27%)	410 (16.02%)	893 (34.88%)	994 (38.83%)

Results (Table 4) show that the PA was higher for the less hateful comments (34.9%, 95% CI = 30.76 – 39.37). The PA for more hateful comments was 7.62% (95% CI = 5.27 – 9.96) which is low compared to less hateful comments.

Results for Gwet's AC2, as with the previous research question, were in concordance with PA. They show that the interrater reliability was higher for less hateful comments compared to more hateful comments.

Figure 3 shows that the 95% confidence intervals do not overlap for both groups, which indicates that the difference in PA between both groups is statistically significant. The previous findings show that raters agree more when they review comments that are less hateful compared to when they review comments that are more hateful.

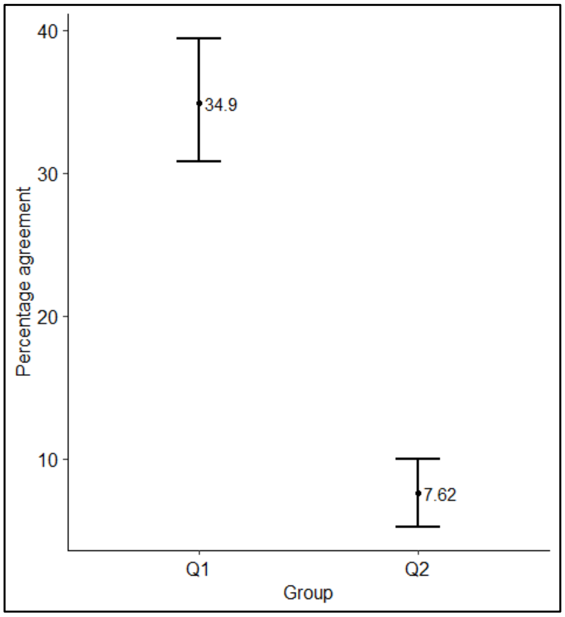

Figure 3: Percentage agreement among less and more hateful comments

5 DISCUSSION

Crowd raters tend to agree most when reviewing comments that are perceived as less hateful. In contrast, the agreement is smaller when the comment is more hateful, and the agreement is lowest for comments with moderate hatefulness. Overall, crowd raters agree more on comments that are less hateful (based on their avg. hatefulness) than on comments that are more hateful (based on their avg. hatefulness).

Thus, the general tendency inclines to be that it is easier for human raters to agree that a comment is either very hateful or not at all hateful, but harder to agree on the middle ground. We explain the results through previous research. Online hatefulness has been found to be a phenomenon that is divisive [4], with difficulties in maintaining and assessing grey areas. Partly, this originates from the subjective nature of online hate [19], and partly due to sarcastic use of language [14] that some individuals can interpret more easily than others. In addition, the interpretation of humor varies, so that a joke that is amusing to one is offensive to another one.

The implication of our findings for researchers developing machine learning models for automatic hate processing are two-fold: (1) *extreme classes are likely to require fewer annotations to reach statistical stability than medium-level hateful comments, because they are more easily agreed upon*; (2) *the variation of hate interpretation among individuals suggests that the machine learning models dealing with online hate should consider empirical distributions of hate ratings rather than averages.* While previous work has done limited advance in this aspect [22], more granular incorporation of hate interpretation distributions is needed, as well as considering individual-level features [20] that impact the hate interpretation of a user.

REFERENCES

[1] 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* (New York, NY, USA, 2015), 1089–1092.

[2] Balbi, S. et al. 2018. Combining different evaluation systems on social media for measuring user satisfaction. *Information Processing & Management.* 54, 4 (2018), 674–685.

[3] Chatzakou, D. et al. 2017. Hate is Not Binary: Studying Abusive Behavior of #GamerGate on Twitter. *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (New York, NY, USA, 2017), 65–74.

[4] Del Vicario, M. et al. 2016. Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific Reports.* 6, (Dec. 2016), 37825. DOI:https://doi.org/10.1038/srep37825.

[5] Efron, B. 1982. *The jackknife, the bootstrap, and other resampling plans.* Siam.

[6] Feng, G.C. 2013. Underlying determinants driving agreement among coders. *Quality & Quantity.* 47, 5 (2013), 2983–2997.

[7] 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 (May 2008), 29–48. DOI:https://doi.org/10.1348/000711006X126600.

[8] Gwet, K.L. 2011. On the Krippendorff's alpha coefficient. *Manuscript submitted for publication. Retrieved October.* 2, (2011), 2011.

[9] Hosseini, H. et al. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *arXiv:1702.08138 [cs].* (Feb. 2017).

[10] Hosseinmardi, H. et al. 2015. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. *Social Informatics* (Dec. 2015), 49–66.

[11] Kwon, K.H. and Gruzd, A. 2017. Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research.* 27, 4 (Jun. 2017), 991–1010. DOI:https://doi.org/10.1108/IntR-02-2017-0072.

[12] Li, C. et al. 2019. Community detection using hierarchical clustering based on edge-weighted similarity in cloud environment. *Information Processing & Management.* 56, 1 (2019), 91–109.

[13] Mohan, S. et al. 2017. The Impact of Toxic Language on the Health of Reddit Communities. *SpringerLink* (May 2017), 51–56.

[14] Rajadesingan, A. et al. 2015. Sarcasm detection on twitter: A behavioral modeling approach. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (2015), 97–106.

[15] Saleem, H.M. et al. 2017. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. *arXiv:1709.10159 [cs].* (Sep. 2017).

[16] Salminen, J. et al. 2018. 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)* (San Francisco, California, USA, Jun. 2018).

[17] Salminen, J. et al. 2018. Inter-rater agreement for social computing studies. *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018)* (Valencia, Spain, Oct. 2018).

[18] Salminen, J. et al. 2018. Neural Network Hate Deletion: Developing a Machine Learning Model to Eliminate Hate from Online Comments. *Lecture Notes in Computer Science (LNCS 11193)* (St. Petersburg, Russia, Oct. 2018).

[19] Salminen, J. et al. 2018. 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)* (Valencia, Spain, Oct. 2018).

[20] Sánchez, P. and Bellogín, A. 2019. Building user profiles based on sequences for content and collaborative filtering. *Information Processing & Management.* 56, 1 (2019), 192–211.

[21] Sood, S.O. et al. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology.* 63, 2 (Feb. 2012), 270–285.

[22] Wulczyn, E. et al. 2017. Ex Machina: Personal Attacks Seen at Scale. *Proceedings of the 26th International Conference on World Wide Web* (Republic and Canton of Geneva, Switzerland, 2017), 1391–1399.