Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings

Joni Salminen[1], Fabio Veronesi[2], Hind Almerekhi[3], Soon-gyo Jung[1] and Bernard J. Jansen[1]
[1] *Qatar Computing Research Institute*
Hamad Bin Khalifa University Doha, Qatar
Email: {jsalminen,sjung}@hbku.edu.qa, jjansen@acm.org
[2] Crop and Environment Sciences Harper Adams Universit Newport, United Kingdom
Email: f.veronesi@gmail.com
[3] The College of Science and Engineering
Hamad Bin Khalifa University Doha, Qatar
Email: halmerekhi@qf.org.qa

***Abstract*— Hate is prevalent in online social media. This has resulted in a considerable amount of research in detecting and scoring it. Most computational efforts involve machine learning with crowdsourced ratings as training data. A prominent example of this is the Perspective API, a tool by Google to score toxicity of online comments. However, a major issue in the existing approaches is the lack of consideration for the subjective nature of online hate. While there is research that shows the intensity of hate varies and the hate depends on the context, there is no research that systematically investigates how hate interpretation varies by country or individual. In this exploratory research, we undertake this challenge. We sample crowd workers from 50 countries, have them score the same social media comments for toxicity and then evaluate the differences in the scores, altogether 18,125 ratings. We find that the interpretation score differences among countries are highly significant. However, the hate interpretations vary more by the individual raters than by countries. These findings suggest that hate scoring systems should consider user-level features when scoring and automating the processing of online hate.**

Keywords— Online hate, social media, hate interpretation

I. INTRODUCTION

Many authors and studies have remarked on the prevalence of online hate in social media [1]–[5], including discussion forums such as Reddit, and social networks such as Twitter, Facebook, and YouTube. The high prevalence of hateful comments has resulted in a considerable volume of research in detecting, classifying, and scoring hateful comments and communities [6]. Most typically, these efforts involve machine learning [7] with crowdsourced ratings as training data [8]. A prominent example of such an approach is the Perspective API, a tool by Google, that scores toxicity of online comments and which has been trained using crowd annotators [9]. However, the major issue in the existing approaches is the lack of consideration for subjective nature of online hate, meaning that they return *one overall score* of online hate that does not consider the interpretation of the user experiencing the hateful comment. Even though there is research that shows the hate interpretation depends on the context [10], such as community standards and special use of language [11], there is no research that would systematically investigate how online hate interpretation varies by country or individual.

In this research, we undertake such a goal. We ask the following research questions:

- Does the country of the rater significantly affect the social media comment hatefulness ratings? *We answer this question by statistically analyzing 18,125 ratings from crowd raters in 50 countries.*
- Is the continent of the rater significantly affecting the social media comment hatefulness ratings? *We answer this question by statistically analyzing 18,125 ratings from 50 crowd workers in countries across multiple continents.*
- Which countries are most sensitive to hate? *We analyze the share of 'very hateful' from all the ratings by country*.
- Which countries are least sensitive to hate? *We answer this question by analyzing the share of 'not hateful at all' ratings from all the ratings by country.*

Overall, we want to know both the sensitivity of people from a given country to hateful online comments, as well as the internal variation of the interpretations of people from that country. We collect hate ratings from crowd workers in 50 countries by showing them hateful social media comments, asking them to rate them for hatefulness and then analyzing the differences in the ratings. To examine the variation of hate interpretation, we define a new metric called *hate interpretation score*.

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

II. Related Literature

Online hate taking place in social media platforms, such as Twitter, Facebook, Reddit, and YouTube has been studied widely in computer science [1], [5], [6], [11]–[13]. However, there is no clear definition that includes all forms of hate. Prior research has shown that online hate has several subcategories like (a) online hate speech, which refers to offensive content that targets a specific group of people [1], [13]; (b) online harassment, where a user deliberately attacks others in an online environment [14], [15]; and (c) cyberbullying, where victims are targeted and intimidated by online social media posts or comments [16]–[18]. Combining the different aspects yields a definition of online hate (often referred to as 'toxicity') that is defined by [19] as degrading remarks towards people where they are humiliated, oppressed, or deemed worthless. Even though this definition implicitly refers to subjective experience in the sense the negative effects of online hate are interpreted by the individuals in their recipient end, the prior research has largely neglected the aspect of the subjective experience. However, we postulate that *what is perceived by degrading may vary greatly by community, country, or individual being exposed to hateful messaging*.

There is some preliminary evidence about the relative nature of online hate. For example, [12] analyzed the health and toxicity of 180 subreddits, finding that there is a high negative correlation between community health and toxicity. However, the researchers found that the toxicity level in each subreddit is contextually dependent on the topic [12]. [20] studied Reddit's ban of *r/CoonTown* and *r/fatpeoplehate* and its impact on the affected users and found that the ban was successful in preventing hate groups from spreading their toxicity among other subreddit communities. [11] employed a community-specific model of hateful speech to avoid troubles that are caused by mere keyword searching, making sure that chosen hate speech communities contain linguistic practices that are different and distinct from other communities. These studies suggest that *the linguistic practices of online hate differ by community*.

The findings by [20] also suggest that relying on self-moderation may not be effective in cases where deviant users may spread toxicity to other communities, calling for automated methods for hate detection. Different approaches have been devised by researchers to detect hateful social media comments, such as dictionary-based [8], distributional semantics [2], multi-feature [5], [21] and neural networks [22], [23]. Yet, these approaches have several shortcomings: they often need to be continually updated, fail at detecting the different degree of hatefulness, are often ineffective against manipulation, and restrict the freedom of users in online communities [24]. For example, users may adjust their wordings (e.g. writing 'joos' instead of 'jews') to fool the algorithms, which hinders language-based detectors [25]. Therefore, the *task of detecting toxicity in online social media comments is challenging due to the difficulty of the language used in such online platforms*.

[8] collected 1,655,131 comments from Yahoo! Buzz, a news website, labeled 6,500 of them for insults and profanity, along with topics of hate: (1) world, (2) business, (3) entertainment, (4) politics, 5) news, and 6) general, and the target was either an author of a previous comment, or a third party. [2] focused on detection of hate speech in user comments collected from Yahoo Finance website. In their context, most insults were targeting rich people [2]. [6] identified several explicit hate targets, e.g. minority groups and women, in Twitter and Whisper, the most common categories for both social media platforms being race, behavior and physical. [11] pulled data from Reddit and Voat and identified three targets of hate speech: Black people, plus-sized people, and women. [5] investigated the social media comments of a major online news media channel and found the most hateful comments targeted the police and media. The conclusion from this line of works is that *there can be several potential targets for online hate, which vary by context*.

There can be seen two challenges in prior research on online hate. Coarse characterizations of hate, e.g. binary models (e.g., hateful/non-hateful) [1]. However, online hate has varying level intensity; that is, a comment is not experienced the same by different online users. Second, the subjective interpretation of online hate is systematically ignored in the models. Even those based on several raters choose an average score for the model, thus producing an average interpretation, usually derived from a small number of raters. Such models may not be representative and, more importantly, they are not adaptive to individual users. In fact, these subjective differences may even be explained as "errors," such as by [1] who noted that humans classify some comments as hateful, even though they were not, according to the consensus that hate speech targets specific groups in a potentially harmful way. Comparing a formal definition to users' subjective experiences is thus challenging. In this research, our goal is to examine the variation of hate interpretation to establish empirical evidence on the subjective experience of online hate, i.e. hate interpretation, and thus show the need for user modeling as an important cornerstone for computational techniques dealing with detection, classification, and moderation of online hate in social media.

III. Method

The data collection for this research had three steps: (1) first, we retrieve a large dataset of comments from social media, namely the YouTube and Facebook channels of a large online news media channel. Second, (2) we filter hateful video comments among the collected dataset. Third, (3) we randomly sample a subset of hateful comments for crowdsourced ratings from people over the world to determine how the interpretation of hate intensity varies. In the following paragraphs, these steps are described in more detail.

A. Research Context

We collect social media comments from a major online news and media company that has an international audience representing viewers from over 200 countries. This media organization has channels on several social media platforms. For this research, we retrieve all the comments from two channels, Facebook and YouTube. By using the application programming interfaces (APIs), we pull comments from videos posted on YouTube and Facebook, including 1,342,597

comments from the period from December 2013 to January 2018. The comments originate from all of the Facebook posts and YouTube videos of the two channels in the two platforms. In the dataset, 24% of the comments are from Facebook, 76% from YouTube. While exploring the comments, we observed that many of them are hateful, i.e., using derogatory language that aims at targeting other people or groups of people.

To elicit hateful comments, we use the technique of open coding [26] to build a hateful phrases dictionary specific to this dataset. This process involves reading the comments and noting down the phrases that are frequently used in a hateful sense. The choice of open coding is supported by prior research showing that the use of hateful expressions has been found to vary by the community in which the commenting takes place [11], [27]. Therefore, it is highly purposeful to build a contextual dictionary to be able to distinguish hateful comments from non-hateful ones. In our context, the topics of the videos are diverse, but most hate, based on our open coding, revolves around political topics (e.g., 'Israel-Palestine', 'Police brutality', 'US presidential elections', etc.). That is why we must adjust the dictionary using open coding, i.e. manually finding hateful expressions that are typical for this dataset. Table I includes examples of hateful words and their appearance in the comments. Overall, the dictionary contains 203 hateful key phrases. We also share this dictionary on GitHub[1] for other researchers to make use of.

TABLE I. EXAMPLES FROM THE CREATED HATE DICTIONARY FOR ONLINE NEWS MEDIA.

Example keyphrase	***Example comment (hateful expressions bolded)***
stupid	Because Denmark is getting smart and Sweden is still **stupid**.
should be killed	however the rapist's **should be killed**.... lets give them a bit of their sharia law they so long for. the **fucking pigs**...
fuck	All the libs making a big deal of it. When in reality if it was Hillary's **bitch ass**, you cowards would be almost certain you had been shot. Right away blaming the 2nd amendment. Soft **pussy fucks**.
zionist	The **zionist scumbags** are losing the battle. #FreePalestine

Second, we apply this dictionary to social media datasets to retrieve hateful comments. Keyword matching and word-based features have been utilized in a similar fashion in several previous works to detect hateful comments [8]. For example, [5] found that simple term-level features were among the best predictors when detecting hateful online comments. After retrieving the hateful comments using the phrases in the dictionary, we find 301,153 hateful comments, representing 22.4% of comments in the whole dataset. It seems therefore evident that the issue of online hate represents a real problem touching online news organizations operating in social media.

[1] https://github.com/joolsa/hatewords

B. Data Validation

To validate that the comments are indeed hateful, we conduct a crowdsourcing experiment, in which we randomly sample 300 hateful comments and 300 non-hateful comments (i.e., in which the hateful key phrases were not present). We then ask crowd raters to evaluate if the comments are hateful or not, defining hateful comments as follows: "*Hateful comment is something that aims to hurt another person or contains bad language*." The results are shown in Table II.

TABLE II. VALIDATION OF DATA USING CROWDSOURCING.

	Hateful (observed)	***Non-hateful (observed)***
Hateful comments (*expected*)	84.47%	15.53%
Non-hateful comments (expected)	19.80%	80.20%

The sanity check yields 84.47% of the hateful comments as hateful and 80.20% of the non-hateful as non-hateful. The observed sentiment of the comments is thus in line with the expected sentiment, and we can proceed with the data collection of interpretation.

C. Collecting online hate interpretations

Finally, we collect the dataset for hate interpretation by asking crowd workers to rate the intensity of hatefulness of the sampled comments. For this, we create a data collection of several data collection tasks in CrowdFlower (a.k.a. FigureEight), the crowdsourcing platform used by several researchers [28]–[30]. We create several tasks in order to mitigate time zone differences in participation and therefore to attract a pool of participants from many countries. Figure 1 illustrates the created tasks.

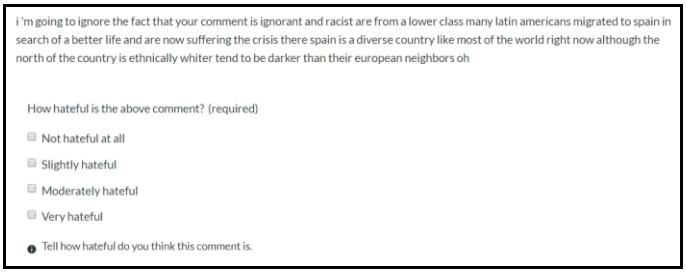

Fig. 1. Question Asked from Crowd Labelers on Hatefulness of the Comments.

We sample the previously retrieved hateful comments and randomly chose 600 comments to be labeled by crowd workers. We instruct the crowd workers to rate if the comment is hateful or not. As in the previous validation task, hatefulness was defined as follows: "*Hateful comment is something that aims to hurt another person or contains bad language*." Crowd workers rate the comments using a 4-point scale so that each comment is 'Very hateful' (4), 'Moderately hateful' (3), 'Slightly hateful' (2), or 'Not hateful at all' (1).

The crowd workers are not systematically sampled by country, but by using the crowdsourcing platform's default settings, we give anyone the possibility to participate. Crowdsourcing, due to the anonymity of participants, often relies on the self-selection of volunteers [31]. This results in the fact that this research is exploratory in nature and describes the sentiment [32]of crowd workers from various countries that were interested in rating online hate at the time this research was conducted. Overall, we have participants from 65 countries and with our budget constraint, obtain 18,125 ratings. To ensure the quality of crowd workers, defined as giving honest opinions, we undertook the following measures:

- **Max Judgments / Rater**: We chose 50, so that with e.g. 300 x 5 = 1500 ratings, there is 50/1500 = 3.33% maximum impact per rater, mitigating individual bias.
- **Minimum time it should take a rater to complete a page of work:** 25 seconds; a page has five comments and we expect it would take a rater at least five seconds to read and rate each (5 x 5 = 25 seconds).
- **Quality Level:** Level 2 (Higher Quality), consisting of a smaller group of more experienced, higher accuracy raters.
- **Judgments per Row:** 5, to get more information on the preferences, as the task is subjective in nature.
- **Price per Judgment:** 4 dollar cents, a 33% increase of the price suggested by the crowdsourcing platform.

Average obtained trust score of the raters was 0.92. The trust score, the maximum value is 1.0, is calculated from the historical accuracy of a crowd worker from all the tasks he or she has taken in the crowdsourcing platform. Here, we rely on this historical score as a metric assessing the crowd worker's honesty as the hate interpretations themselves are subjective and we are thus unable to define a truthful baseline. Historical accuracy has been found useful when classification tasks for the crowd involve subjectivity [33].

D. Data processing

Prior to performing statistical analyses, we perform the necessary data manipulations. First, we create a contingency table with a count of hatefulness ratings per country. We compute the number of records per country to understand whether to exclude some countries based on low sample sizes. Figure 2 illustrates the number of participants from each country in the dataset.

After considering various rules of thumb for narrowing down the data, we decided to exclude all countries with less than 65 records (value of Canada). This leaves us 50 countries out of 65 total countries, involving countries from several continents, including Africa, North-America, Europe, and Asia. Because there is an uneven number of observations per country, we decide to work with proportions in the statistical analyses. Figure 2 illustrates the number of datapoints from different countries.

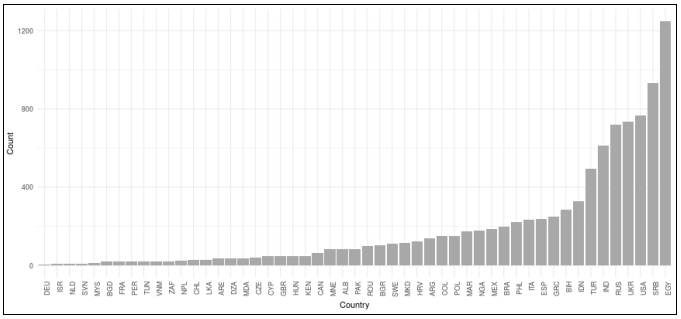

Fig. 2. Number of participants from different countries.

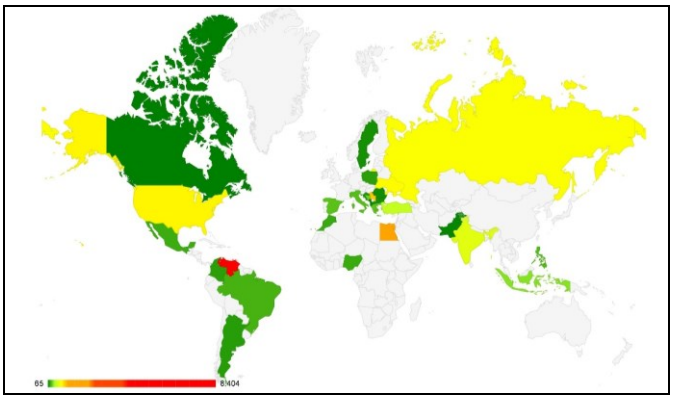

Fig. 3. Number of participants from different countries.

IV. FINDINGS

A. Country-level differences

Since we are working with proportions, we fit a generalized linear model (GLM) with a binomial distribution, in which we predict the hatefulness rating using the Country variable. Results from the likelihood ratio test are shown in Table 2. From Table 2, we can see that the Country variable is highly significant. By looking at the numerical results, we answer the questions on most/least sensitive countries to hate. The countries in Table 3 have the highest proportions of answers in the categories "Very hateful" and "Not hateful at all".

TABLE III. EXAMPLES FROM THE CREATED HATE DICTIONARY FOR ONLINE NEWS MEDIA.

	LR Chisq	Df	Pr(>Chisq)
How_Hateful	58.10	5	<0.0001
Country	580.84	29	<0.0001
How_Hateful:Country	1751.93	145	<0.0001
	LR Chisq	Df	Pr(>Chisq)

TABLE IV. MARGINAL PREDICTED MEAN PROBABILITIES FROM GLM. MOST SENSITIVE TO HATE ARE FROM 'VERY HATEFUL' CATEGORY AND LEAST SENSITIVE FROM 'NOT HATEFUL AT ALL' CATEGORY.

Most sensitive	Prob.	Least sensitive	Prob.
Pakistan	0.264	Albania	0.254
Spain	0.194	Bosnia and Herzegovina	0.246
Poland	0.174	Canada	0.244
Makedonia	0.170	Morocco	0.224
Greece	0.169	United States	0.214

As we further investigate the dataset, we notice that the variance within countries is very high. This means there are large variations between answers from people living in the same country. That does not mean there are no differences between countries, but it simply means that these are smaller than variation within countries. Within-country variation will be explored in association with hate interpretation score.

B. *Continent level differences*

For modeling continents, we first need to import a table with the continent for each country. We can do this by importing a country-continent table available on Wikipedia[2]. This table includes a list of countries associated with continent information (e.g., AF – Africa, AS – Asia, etc.). We match the value of the continent column of the table from Wikipedia with the name of the country in the dataset. On Wikipedia, Turkey appears twice, once in Europe and once in Asia; we define the country to belong to Asia using a customized function. At this point, we can proceed with the same modeling as done before (i.e. GLM with binomial distribution), using Continent as a covariate instead of Country. Table V shows the results of the likelihood ratio test.

TABLE V. EFFECT OF CONTINENT ON HATEFULNESS.

	LR Chisq	Df	Pr(>Chisq)
How_Hateful	338.53	5	<0.0001
Continent	343.36	4	<0.0001
How_Hateful:Continent	703.39	20	<0.0001

C. *Hate interpretation scores*

For the next step of the analysis, we introduce the hate interpretation score (HIS). This score simply represents the deviation of the aggregated sum of raters from a given country from the average of all ratings. For each comment, there are *n* ratings of hatefulness. First, we compute the average hatefulness rating for each comment based on these *n* ratings. Then, we compute the difference for each unique country from the average rating. For example, if the average hate rating of Comment_1 is 3.5 and Mexican raters would rate it 2, then the difference from the average is -1.5. We perform this computation for all comments and then take the mean of each country's differences, which is the HIS of that country.

[2] https://en.wikipedia.org/wiki/List_of_sovereign_states_and_dependent_territories_by_continent_(data_file)

To calculate the HIS, we run a loop that iterates through the unique identifiers for the comments. Then, for each comment, it computes the average score, and the deviation between each rating and the average. Results are then aggregated by country and written to a file for statistical analysis. To remove the effect of an unbalanced number of observations per country, we scale and center the values so that they range from -1 to +1. Figure 4 illustrates the scores for each country.

To statistically compare the HIS' between the countries, we first check that the interpretation score is normally distributed. The skewness of this distribution of HIS' is 0.019, which is extremely low and indicates normally distributed data. Normality enables us to run a linear model, the results are shown in Table 6.

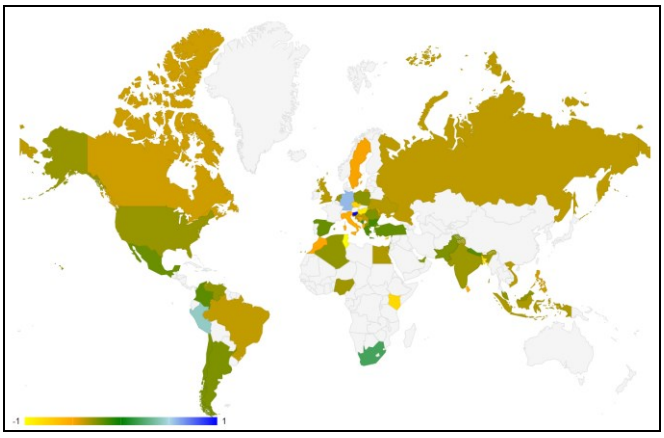

Fig. 4. Normalized hate interpretation scores at the country level.

TABLE VI. ANALYSIS OF VARIANCE FOR HIS BETWEEN COUNTRIES.

	Df	Sum Sq	Mean Sq
Country	29	62.3	2.14983
Residuals	17495	14613.1	0.83527
F value	Pr(>F)		
2.5738	<0.0001		

From Table 6 it is easy to observe that the residuals sum of squares, also known as the within groups sum of squares, is much larger than the between groups sum of squares. This implies that there is a large variation in the hate interpretation scores within countries than between them. To further highlight this point, we compute the effect size, which shows the effective magnitude of differences between countries. We compute the effect size as the partial Eta-Squared. The resulting partial Eta-Squared is extremely small (pETAsq = 0.0042), indicating once again that the variation within groups is much larger than the variation between them.

Running a multiple comparison model, we find, once again, that despite the variable Country being highly significant for the ANOVA model, there are many countries for which the multiple comparison cannot find statistically significant differences. To compare the countries, we decide to use the standard error of the mean as a measure of variance, since it provides a way to assess the statistical differences between the

HIS'. The standard error tells the level of reliability of the mean value represented in the hate scores. We compute the standard errors of the mean and plot a bar-chart with error bars representing these values (see Figure 5).

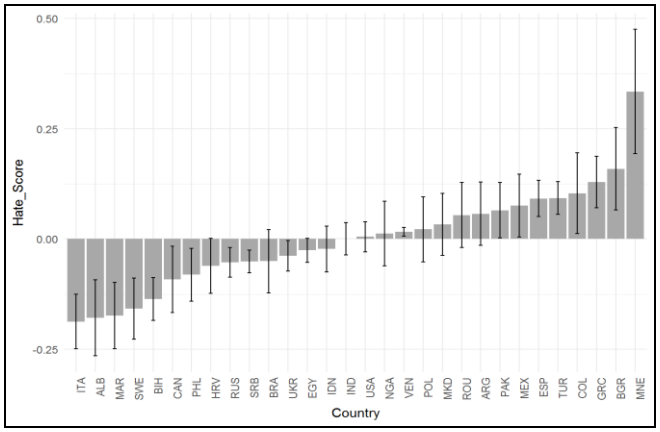

Fig. 5. Normalized hate interpretation scores at the country level.

From Figure 6, we observe that the variance between countries is lower compared to the variance within countries. Figure 4 further illustrates this point, showing four the variance of HIS for four countries (United States, Italy, Columbia, and Russia). Similar variational differences can be found across the set of countries.

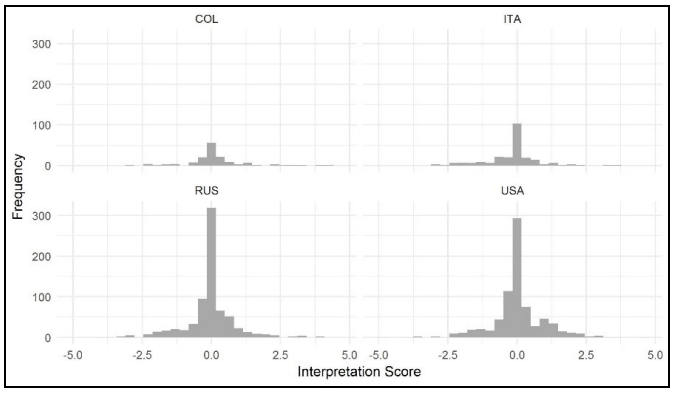

Fig. 6. Normalized hate interpretation scores at the country level.

V. Discussion and Implications

Empirically, our main finding is that hate interpretation differs more by individuals than by countries, although on average, hate interpretation differs significantly between the countries in the sample. Conceptually, our findings of this variation in hate interpretation lend support to the argument that *online hate should be defined as a subjective experience* rather than as an average score that is uniform to all users, as done by the existing body of literature focused on online hate detection and classification [2], [5], [6], [13].

Therefore, the findings have implications for the computational online hate research currently under tremendous interest in the academia. The models developed in the past research exclusively focus on determining *one overall score* for hatefulness of a social media comment, whereas, in reality, the users' subjective experience of the hatefulness of the comments varies. By neglecting the interpretation aspect, the models are missing key information and risk overestimating or underestimating hatefulness for any specific user of the social platform, thus resulting in grave mistakes when applied to automated or semi-automated moderation systems. For example, deleting, hiding, or recommending comments should be done considering the interpretation of a particular user, rather than the aggregate interpretation of all users. Therefore, our argument is to incorporate user-level features into modeling of online hate, currently not done in any of the research we are aware of. As a stepping stone toward that direction, this exploratory analysis has shown empirical evidence of the different hate interpretations by individuals between and within a country. Those differences should be measured in greater detail, considered in model development, and then implemented in a system to measure the impact on the user experience of online social media users.

The study also involves some room for improvement. From a statistical point of view, the most notable limitation is that we did not conduct systematic sampling but used a form of convenience sampling based on the willingness of people from different countries to participate in the rating of online hate. Therefore, to substantiate and expand the findings, future research should systematically sample different countries and continents. Another limitation is that, due to the nature of crowd workers being anonymous, we are missing important background variables, such as age, gender, occupation, education level, political ideas and other variables that potentially influence hate interpretation. As said, due to the anonymous nature of the crowd [34], these variables were not available for this study. However, future research could utilize different means for data collection, to better understand what drives the perceptual differences of online hate. The study at hand provides evidence that such differences do exist.

References

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in *Proceedings of Eleventh International AAAI Conference on Web and Social Media*, Québec, Canada, 2017.

[2] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate Speech Detection with Comment Embeddings," in *Proceedings of the 24th International Conference on World Wide Web*, New York, NY, USA, 2015, pp. 29–30.

[3] J. Hawdon, A. Oksanen, and P. Räsänen, "Online extremism and online hate: Exposure among adolescents and young adults in four nations," *Nordicom Information: Medie-och kommunikationsforskning i Norden*, vol. 37, no. 3–4, pp. 29–37, 2015.

[4] P. Räsänen, J. Hawdon, E. Holkeri, T. Keipi, M. Näsi, and A. Oksanen, "Targets of online hate: Examining determinants of victimization among young Finnish Facebook users," *Violence and victims*, vol. 31, no. 4, p. 708, 2016.

[5] J. Salminen *et al.*, "Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media," in *Proceeding of The International AAAI Conference on Web and Social Media (ICWSM 2018)*, San Francisco, California, USA, 2018.

[6] M. Mondal, L. A. Silva, and F. Benevenuto, "A Measurement Study of Hate Speech in Social Media," in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, New York, NY, USA, 2017, pp. 85–94.

[7] H. T. Nguyen and M. Le Nguyen, "Multilingual opinion mining on YouTube – A convolutional N-gram BiLSTM word embedding," *Information Processing & Management*, vol. 54, no. 3, pp. 451–462, May 2018.

[8] S. O. Sood, E. F. Churchill, and J. Antin, "Automatic identification of personal insults on social news sites," *J. Am. Soc. Inf. Sci.*, vol. 63, no. 2, pp. 270–285, Feb. 2012.

[9] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving Google's Perspective API Built for Detecting Toxic Comments," *arXiv:1702.08138 [cs]*, Feb. 2017.

[10] S. Balbi, M. Misuraca, and G. Scepi, "Combining different evaluation systems on social media for measuring user satisfaction," *Information Processing & Management*, vol. 54, no. 4, pp. 674–685, 2018.

[11] H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths, "A Web of Hate: Tackling Hateful Speech in Online Social Spaces," *arXiv:1709.10159 [cs]*, Sep. 2017.

[12] S. Mohan, A. Guha, M. Harris, F. Popowich, A. Schuster, and C. Priebe, "The Impact of Toxic Language on the Health of Reddit Communities," in *SpringerLink*, 2017, pp. 51–56.

[13] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the Targets of Hate in Online Social Media," in *Proceedings of Tenth International AAAI Conference on Web and Social Media*, Palo Alto, California, 2016.

[14] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial Behavior in Online Discussion Communities," *arXiv:1504.00680 [cs, stat]*, Apr. 2015.

[15] H. Yin, B. Cui, L. Chen, Z. Hu, and Z. Huang, "A Temporal Context-aware Model for User Behavior Modeling in Social Media Systems," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 2014, pp. 1543–1554.

[16] L. Bowler, C. Knobel, and E. Mattern, "From cyberbullying to well-being: A narrative-based participatory approach to values-oriented design for social media," *J Assn Inf Sci Tec*, vol. 66, no. 6, pp. 1274–1293, Jun. 2015.

[17] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network," in *Social Informatics*, 2015, pp. 49–66.

[18] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "Careful What You Share in Six Seconds: Detecting Cyberbullying Instances in Vine," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, New York, NY, USA, 2015, pp. 617–622.

[19] J. Knight, *Better Conversations: Coaching Ourselves and Each Other to Be More Credible, Caring, and Connected*, 1 edition. Thousand Oaks, California: Corwin, 2015.

[20] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You Can'T Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech," *Proc. ACM Hum.-Comput. Interact.*, vol. 1, no. CSCW, pp. 31:1–31:22, Dec. 2017.

[21] 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*, Republic and Canton of Geneva, Switzerland, 2016, pp. 145–153.

[22] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Detecting Offensive Language in Tweets Using Deep Learning," *arXiv preprint arXiv:1801.04433*, 2018.

[23] H. Yenala, A. Jhanwar, M. K. Chinnakotla, and J. Goyal, "Deep learning for detecting inappropriate content in text," *International Journal of Data Science and Analytics*, Dec. 2017.

[24] A. Massanari, "#Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures," *New Media & Society*, vol. 19, no. 3, pp. 329–346, Mar. 2017.

[25] W. Warner and J. Hirschberg, "Detecting Hate Speech on the World Wide Web," in *Proceedings of the Second Workshop on Language in Social Media*, Stroudsburg, PA, USA, 2012, pp. 19–26.

[26] J. M. Corbin and A. Strauss, "Grounded theory research: Procedures, canons, and evaluative criteria," *Qualitative sociology*, vol. 13, no. 1, pp. 3–21, 1990.

[27] K. H. Kwon and A. Gruzd, "Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos," *Internet Research*, vol. 27, no. 4, pp. 991–1010, Jun. 2017.

[28] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating named entities in Twitter data with crowdsourcing," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 80–88.

[29] W. Huang, I. Weber, and S. Vieweg, "Inferring Nationalities of Twitter Users and Studying Inter-National Linking," *ACM HyperText Conference*, 2014.

[30] C. Van Pelt and A. Sorokin, "Designing a scalable crowdsourcing platform," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 765–766.

[31] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic, "An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets.," *ICWSM*, vol. 11, pp. 17–21, 2011.

[32] M. Tubishat, N. Idris, and M. A. Abushariah, "Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges," *Information Processing & Management*, vol. 54, no. 4, pp. 545–563, 2018.

[33] 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*, New York, NY, USA, 2015, pp. 101–110.

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