How do you feel?

Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection

Lammerts, Philippe; Lippmann, Philip; Hsu, Yen Chia; Casati, Fabio; Yang, Jie

DOI
10.1145/3600211.3604655

Publication date
2023

Document Version
Final published version

Published in
AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society

How do you feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection

Philippe Lammerts
philippelammerts@gmail.com
Delft University of Technology
Delft, The Netherlands

Philip Lippmann
p.lippmann@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Yen-Chia Hsu
y.c.hsu@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Fabio Casati
fabio.casati@servicenow.com
ServiceNow
Santa Clara, CA, USA

Jie Yang
j.yang-3@tudelft.nl
Delft University of Technology
Delft, The Netherlands

ABSTRACT

Hate speech moderation remains a challenging task for social media platforms. Human-AI collaborative systems offer the potential to combine the strengths of humans' reliability and the scalability of machine learning to tackle this issue effectively. While methods for task handover in human-AI collaboration exist that consider the costs of incorrect predictions, insufficient attention has been paid to accurately estimating these costs. In this work, we propose a value-sensitive rejection mechanism that automatically rejects machine decisions for human moderation based on users' value perceptions regarding machine decisions. We conduct a crowdsourced survey study with 160 participants to evaluate their perception of correct and incorrect machine decisions in the domain of hate speech detection, as well as occurrences where the system rejects making a prediction. Here, we introduce Magnitude Estimation, an unbounded scale, as the preferred method for measuring user (dis)agreement with machine decisions. Our results show that Magnitude Estimation can provide a reliable measurement of participants' perception of machine decisions. By integrating user-perceived value into human-AI collaboration, we further show that it can guide us in *1)* determining when to accept or reject machine decisions to obtain the optimal total value a model can deliver and *2)* selecting better classification models as compared to the more widely used target of model accuracy.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; **Social media**.

KEYWORDS

value-sensitive machine learning, rejection, machine confidence, crowdsourcing, human-in-the-loop, hate speech

AIES '23, August 08–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604655

ACM Reference Format:
Philippe Lammerts, Philip Lippmann, Yen-Chia Hsu, Fabio Casati, and Jie Yang. 2023. How do you feel? Measuring User-Perceived Value for Rejecting Machine Decisions in Hate Speech Detection. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3600211.3604655

1 INTRODUCTION

Hateful content spread online through social media remains a significant problem. Ignoring its presence can lead to psychological harm and even result in violence and other conflicts [35, 43, 48, 50]. Governmental institutions and social media platforms are increasingly aware of these risks and are combating hate speech. For example, the European Union developed a Code of Conduct on countering hate speech [21], requesting large social media companies to moderate hate speech and report their progress yearly. However, results reported so far are not yet satisfactory, as, for example, less than 5% of hateful content has been removed from Facebook [28].

Hateful content moderation is either carried out manually or automatically by computational algorithms, where manual moderation may be more reliable but is not scalable to handle the deluge of user-generated content [38]. Further, continuous exposure to harmful content can be harmful to moderators as it can induce mental issues and potentially even lead to acts of self-harm [61]. Computational solutions are, therefore, urgently in demand by online platforms [24]. The methods considered best suited to this task are mainly based on machine learning, which has achieved reasonable performance at scale [25]. Yet, machine learning methods are far from being reliable, especially in dealing with hateful content previously unseen in the training data, which is often limited in size and biased [4]. Several recent studies on hate speech have shown a significant drop in machine learning performance when assessed on different data from those captured in the training phase [3, 32].

An approach that can combine the strengths of both previously mentioned approaches is human-AI collaboration, where humans are involved to solve AI-hard tasks, typically by taking over decisions where machines are unreliable [12, 14]. Such an approach is favorable in applications where decisions involve high-stakes and incorrect decisions can lead to damaging effects, as is the case for hate speech detection. Human-AI collaboration has been advocated in the human computation community [14, 53, 68] and, likely, is also an approach widely being used in enterprise applications

such as search and conversational agents [37]. Despite this, methods for implementing human-AI collaboration so far are limited to predefined heuristics and have largely ignored the complexity of real-world problems, especially the cost of incorrect predictions being context-dependent.

Common heuristics of task handover from machines to humans are based on machine confidence: humans take over the task when the confidence of the machine in its decision is lower than a predefined threshold [12]. Such heuristics assume that machine confidence is well-calibrated, that is, a decision with high confidence should be more likely to be reliable and vice versa. This assumption however does not hold for many machine learning models, especially deep learning models, which may indicate high confidence when decisions are incorrect or vice versa [5, 33]. An improved approach is proposed by Geifman and El-Yaniv [26] which determines the appropriate confidence threshold based on empirical evidence of machine correctness, e.g., based on the accuracy-threshold curve obtained on an empirical dataset. Such an approach, however, does not take into account the implications of right or wrong decisions. Incorrect decisions in high-stakes domains have a larger impact that, in turn, should pose a stricter constraint on accepting machine decisions, e.g., via a higher confidence threshold. Similar ideas have recently been discussed in position papers that advocate the adoption of the notion of context-dependent *value* as a replacement of accuracy, the most common metric in machine decisions assessment [13, 60]. Value, however, is an abstract term – it can be interpreted from social, ethical, or commercial perspectives [17, 29, 70] – yet the discussion on what creates value and how to measure it, specifically in a machine learning context, is limited due to it depending on the application.

In this paper, we study the problem of operationalizing value perception of machine decisions and its integration into human-AI collaboration in the specific context of hate speech detection. We start by identifying several factors that may affect the value definition, namely the selection of a specific stakeholder's standpoint and the relativity of value perception as affected by stakeholder expectation or regulation. We then operationalize user-perceived value in hate speech moderation scenarios, where a decision with a corresponding confidence has been made by a machine. To measure these perceptions, we explore several measurement scales and propose to select Magnitude Estimation (ME) [62] as the primary scale. ME allows the measurement of the magnitude of user (dis)agreement using an unbounded scale and makes it possible to obtain the relative ratios between the magnitudes of different machine decisions. These ratios are essential to determine the optimal confidence threshold for rejecting machine decisions (see section 2).

To validate ME in value operationalization, we designed a survey study where we recruited 160 participants. Each participant's perception regarding a dataset of 40 selected hateful and non-hateful tweets and their (dis)agreement regarding the corresponding machine decisions were evaluated. Through a between-subject study, we show that Magnitude Estimation returns results with significantly higher inter-rater reliability compared to other scales, showing its suitability in measuring user perception. Our results show that the inter-rater reliability is significantly higher for incorrect decisions than for correct decisions, indicating a strong consensus among participants regarding the consequences of harm, as well as disagreements on what constitutes hate online. Further, users appear to be more negatively affected when a non-hateful post is subject to moderation than when an instance of hate speech is classified as non-hateful, implying that users would rather contend with an instance of hate speech than have an innocent user punished for a non-hateful post.

To demonstrate the utility of value integration in human-AI collaboration, we evaluate the effect of rejecting machine decisions made by three machine learning-based hate speech detection models – including traditional, deep learning, and BERT-based models [19] – in handling data from both seen and unseen sources. Our results show that for all three models, when evaluated on unseen data, the optimal confidence thresholds determined by the model-delivered value are much higher than the optimal thresholds on seen data. These results confirm the findings from previous studies on machine biases and demonstrate the effectiveness of using value as a target for optimally rejecting machine decisions. We further show that when selecting the optimal model, using value as the criterion returns different results compared to using accuracy. Note, that our approach to measuring value perception can be applied to different tasks and is model-agnostic.

In summary, we make the following key contributions:

- We introduce Magnitude Estimation as a scale for measuring user perception of machine decisions in scenarios where these decisions are correct and incorrect;
- We demonstrate the applicability of Magnitude Estimation through a between-subject survey study, as well as the utility of value for optimally rejecting machine decisions;
- We contribute a set of insights into user-perceived value of automated machine decisions, especially their attitudes towards different types of (mis)classifications.

2 BACKGROUND ON VALUE-SENSITIVE REJECTION OF MACHINE DECISIONS

This section introduces the background of value-sensitive rejection of machine decisions in a hybrid human-AI workflow, based on previous work [59, 60], and subsequently identifies factors that influence value perception in hate speech detection.

2.1 Rejection for Binary Classification

We consider the general case of human-AI collaboration as follows: the machine decision can either be accepted or rejected; if rejected, the decision will be taken over by a human decision maker. Formally, consider a binary classification problem for which we have a machine learning classifier, whose output on a data item x is confidence, $\mathbf{c}$, (e.g., the output from the softmax layer of a neural network). The rejection is dependent on a threshold denoted by $\tau \in [0, 1]$, which then modifies the final output of the machine as

$$\hat{y} = \begin{cases} y, & \mathbf{c}_y \geq \tau, \\ y_r, & \text{otherwise.} \end{cases} \tag{1}$$

where y denotes an accepted decision and y_r denotes the special decision of rejection, resulting in a human making the final decision.

We now discuss how the optimal confidence threshold for rejecting machine decisions is affected by the value formulation. We consider the binary classification case: when the machine decision

is either positive (i.e., the content is deemed hateful) or negative (i.e., non-hateful). There is a value, V, attached to each of these, depending on whether this positive or negative decision is correct or not. This results in true positive (TP), true negative (TN), false positive (FP), false negative (FN), and rejected predictions as possible outcomes. V_{TP} and V_{TN} are positive, while V_{FP}, V_{FP}, and rejected predictions, V_r, are negative (i.e., costs). The optimal threshold for positive classifications is:

$$\tau_O^p = \frac{V_{FP}}{V_{FP} - V_{TP}} = \frac{\gamma^p}{\gamma^p + 1} \tag{2}$$

if we assume $V_{FP} = -\gamma^p \cdot V_{TP}$, that is, the cost of a false positive is γ^p times worse than the value of a true positive. Similarly, in the case of negative classifications, the optimal threshold would be $\tau_O^n = \frac{\gamma^n}{\gamma^n+1}$ where $V_{FN} = -\gamma^n \cdot V_{TN}$, i.e., the cost of false negative is γ^n times worse than the value of a true negative.

When the cost of incorrect decisions is very high, i.e., $\gamma \gg 1$, the optimal confidence threshold would tend close to 1, meaning almost all machine decisions are rejected. When the cost of an incorrect decision is very low, i.e., $\gamma \approx 0$, the optimal threshold would be close to 0, and virtually all machine decisions are accepted. These results, therefore, follow our intuition. An important conclusion we can draw from equation (2) is that the optimal threshold is dependent *only on the ratio* of the value (or cost) between an incorrect decision and that of a correct one (per class).

Threshold optimization is the process of finding the threshold that maximizes value empirically. If a system is calibrated before use, simulations can be used to find the optimal theoretical threshold, which is the optimal τ that maximizes value. In this paper, τ is determined by means of calibration, done by means of temperature scaling [47], followed by a calculation of the theoretical threshold based on the crowdsourced survey data, as it allows us to quantify and compare the opinions of participants on the value of true and false predictions and thus compute the ratios for our use case.

2.2 Value Factors in Hate Speech Detection

We denote the value of classifying a data item correctly, or incorrectly, and that of rejecting a classification as V_c, V_w, and V_r, respectively. We make the following observations when considering value for hate speech detection: 1) Value is dependent not only on the machine learning model but also on the specific context to which the model is applied. For example, an incorrect prediction in the medical domain potentially has a bigger impact than one in e-commerce. In a high-stakes domain, generally, we would assume $V_c > V_r > V_w$ and thus a correct machine decision saves the cost of human moderation and accelerates the decision-making process, while a rejection requires additional human intervention. 2) Value interpretations from different stakeholders can vary. In hate speech detection, for example, a rejection of a machine decision induces the cost of human moderation from the business perspective, while from the user perspective what is more important is the exposure to hateful content. In our study, we choose to take the user's standpoint, and, as such, view V_r to come with an inherent cost since human moderation will be pending and the potentially hateful content will remain visible. 3) Value is affected by both stakeholder expectations and regulation. For example, in the hate speech detection case, when hateful content is posted, from the user's perspective, the value derived from a correct machine decision depends on the user's general expectation of how hateful content should be handled. Similarly, the legality of hate speech in certain jurisdictions may influence stakeholder perception.

Given the above observations, we now introduce the function to determine the total value, $V(\tau)$, of a given model with a reject option at the rejection threshold τ on a given dataset. Assuming that when accepted, correct decisions increase the overall value and when rejected, they decrease the overall value and vice versa, then, $V(\tau)$ may be formalized as:

$$V(\tau) = \sum_p (V_p - V_r)N_p + \sum_q (V_r - V_q)N_q, \tag{3}$$

where $p \in [TP, TN, FP, FN]$, $q \in [TP, TN, FP, FN]$, and N_p and N_q are the number of accepted and rejected data items for the difference scenarios, respectively. Note, that we assume that rejected decisions have a cost that decreases the overall value, i.e., V_r is negative, as users have to wait on a moderation decision. Thus, equation (3) allows us to summarize the value gained and the cost subtracted into a single value for the model by considering the value or cost of each scenario and how often it occurs, while also taking the cost of rejection into account.

3 SURVEY STUDY

To define the relative value of scenarios, we design a survey to ask participants the degree to which they agree or disagree with the decisions of a fictional social media platform, SocialNet. These scenarios represent TP, TN, FP, FN, and rejected predictions. The TP and TN scenarios imply that SocialNet successfully detects whether a post is hateful or not hateful, respectively. The FP scenario means that SocialNet incorrectly predicts a non-hateful post as hateful, and conversely for the FN scenario. For example, in the FN scenario, the survey shows a hateful post to the subject and explains that SocialNet did not identify the post as hate speech.

3.1 Choice of the Scale

We use ME as the primary scale. A Likert scale was initially considered, as it is widely used in research for retrieving participant opinions and is perhaps more intuitive for participants [10]. However, a Likert scale is not suitable in our case, as Likert-type items are ordinal, meaning that we only know the ranks but not the exact distances between the items [2]. In our case, computing the relative values (i.e., ratios) of our scenarios requires measuring the distances between different items, which cannot be provided by a Likert scale. On the contrary, the ME scale allows us to measure ratios by asking participants to provide numerical ratings. ME originated from psychophysics, where participants gave quantitative estimates of sensory magnitudes [62]. For sound loudness, a sound twice as loud as the previous one, should ideally receive a rating twice as large.

Researchers have previously applied the ME scale to different physical stimuli (e.g. line length, brightness, or duration) and proved that the results are reproducible, as well as that the data has ratio properties [46]. Other works have shown that the ME technique is also helpful for rating abstract types of stimuli, such as judging the relevance of documents [42], the linguistic acceptability of

sentences [7], and the usability of system interfaces [45]. Thus, we conclude that ME is a promising method for judging hate speech.

3.2 Normalization and Validation of the Scale

The ME scale is unbounded. For example, suppose we first show a scenario and the participant provides a value (e.g., 100) to indicate the degree of agreement. Suppose we next present a scenario that the participant agrees with more. The participant can always provide a higher value (e.g., 125) and not be restricted within a fixed range. The results need to be normalized as different participants rate the agreement/disagreement degree differently.

Multiple solutions exist for normalizing the ME scale, such as modulus normalization, which uses geometric averaging to preserve the ratio information [45, 46]. Unlike the unipolar ME scales used in previous research [7, 45], we use bipolar scales. Using arithmetic averaging is inappropriate since it uses logarithmic calculations and would disrupt the ratio scale properties [46]. Therefore, we normalize the results by dividing the magnitude estimates of each subject by their maximum estimate. We multiply the normalized magnitude estimates by 100 for the sake of clarity. This way, all magnitudes estimates are in the range $[-100, 100]$ while maintaining their ratio properties.

Most previous research using the ME scale applies validation, such as cross-modality validation, where estimated magnitudes are compared to the physical stimuli using correlation analysis [7]. Cross-modality validation is difficult in domains that do not have exact measures of stimuli, such as hate speech. Some previous work compared ME with other validated scales [42]. In our case, we use the 100-level scale to validate the ME scale by analyzing their correlation [57], which is a form of convergent validation [22].

3.3 Participants and Data

We use Prolific to recruit crowd workers for the study.[1] Participants need to be at least 18 years of age, be fluent in English, and have an approval rating of over 90%. Participants also need to have experience using a social media platform regularly (at least once a month). Every participant is paid an hourly wage of 9 GBP, exceeding the UK minimum wage at the time of the study. Regarding sample size, we recruit 24 participants for the pilot study and 136 participants for the official study. Of the recruited participants, 50% identified as female, though Gold and Zesch [30] showed that there is no significant difference when perceiving hate between genders. Half of the participants are assigned the ME scale and the other half the 100-level scale. We choose a 90% Confidence Interval (CI) and 10% Margin of Error (MoE) for this study due to budget limitations. There are billions of social media users, and according to Müller et al. [49], we need a sample size of 68 participants per measurement scale, i.e., 136 participants, to reach the desired CI and MoE.

The final dataset consists of 20 hateful and 20 non-hateful social media posts from a public dataset [8] to build the machine decision scenarios (TP, TN, FP, FN, and rejection). The dataset contains 13,000 English tweets, and each tweet is annotated with three categories: hate speech (yes/no), target (group/individual), and aggressiveness (yes/no). We first exclude tweets that are replies or contained mentions or URLs since they have unclear contexts. Finally, we use clustering analysis to select 40 tweets for our study. We use a cluster size of 20 for the non-hateful tweets and sample one tweet per cluster by taking the nearest sample to each cluster centroid to obtain each cluster's most representative tweets. For the hateful tweets, we first divide them into four groups using the target and aggressiveness categories. Similarly, for each hateful tweet group, we use a cluster size of 5 and sample one tweet per cluster. We perform latent semantic analysis (LSA), which is a combination of term frequency-inverse document frequency (TF-IDF) and Singular Value Decomposition (SVD), and k-means clustering on each group of tweets. We calculate the silhouette coefficient to determine the optimal cluster size (k value) for the neutral tweets and the four groups of hateful tweets. We manually select one tweet per cluster using a majority vote from three members of our group to choose representative tweets and create the final set of 40 tweets.

Additional information on the study's variables, pilot study, demographics, as well as example tasks may be found in appendix A.

3.4 Procedure and Data Quality Control

The survey first presents the informed consent policy and excludes participants that do not agree with it. Next, introductory texts are shown to explain the possible machine decisions. In the case of using the ME scale, participants are presented with a warm-up task to estimate different line lengths. Then, the survey asks 40 randomly shuffled question sets regarding the TP, TN, FP, FN, and rejection scenarios (with 8 question sets per scenario). The first question is about whether participants think the post is hateful (yes/no). The second question is whether participants agree or disagree with the decision made by the machine, which may be correct or incorrect, or are neutral towards it. In the case of a non-neutral decision, the survey asks the third question about the degree to which participants agree or disagree with the machine's decisions, using either the ME or 100-level scale, depending on their group. There is no time limit for the survey.

In the middle of the question sets, we use two Instructional Manipulation Checks to determine if the user is paying attention[2]. These attention checks ask participants to select a specific option from multiple choices (e.g., "You must select Orange"). We exclude responses from the participants who fail the attention checks or do not complete all questions. For the ME scale, we discard responses that do not perform well in the line length warm-up task.

3.5 Analysis

We first compute the values for the TP, TN, FP, FN, and rejection scenarios using the survey study data. For both scales, we convert disagreement (with the machine decision) ratings to negative values, neutral stances to 0, and agreement ratings to positive values. We apply convergent validity, in which a correlation analysis between different scales (i.e., the ME and 100-level scales) is conducted to determine if they measure the same phenomenon [22]. We expect a medium-large correlation between both scales, meaning that ME responses small in magnitude should correspond to 100-level scale responses small in magnitude and vice versa. Finally, we analyze reliability, which determines whether we can trust our results and

[1] Approved by the ethics committee of our organization.

[2] Prolific's Attention and Comprehension Check Policy

	ME		S100	
	α	v	α	v
TP	0.07	18.15	0.04	77.00
TN	0.10	36.32	0.11	86.31
FP	0.39	-16.69	0.07	-51.00
FN	0.92	-28.08	0.14	-62.43
Rejection	-0.31	-4.82	0.07	-16.37
All	0.78	—	0.44	—

Table 1: Krippendorff's alpha (α) and the scenario values (v) for TP, TN, FP, FN, and rejection scenarios. ME refers to Magnitude Estimation, and S100 refers to the 100-level scale.

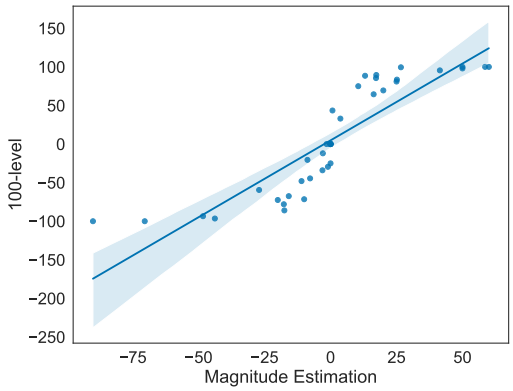


Figure 1: Correlation plot between the median normalized magnitude estimates and the median 100-level scores per question, showing agreement and disagreement.

achieve consistent outcomes [22]. In our case, we use inter-rater reliability to investigate whether different subjects give approximately the same judgments to the same scenarios and, thus, whether the degree to which hate speech is subjective. It is measured using Krippendorff's alpha, which we calculate using the normalized ME and 100-level values for all scenarios.

4 RESULTS

4.1 Reliability and Validity

First, for each survey question set, we calculate the median of all responses. This step yields 40 values (eight values per scenario). We use the median since data from both scales are highly skewed. Then, we calculate the mean of the values ($V_{TP}, V_{TN}, V_{FP}, V_{FN}, V_r$) within each scenario, giving us the final five values for the TP, TN, FP, FN, and rejection cases. The results for both scales can be seen in table 1. The total value, V, is calculated at a later point in this section using the different values.

We calculate Krippendorff's alpha to measure the inter-rater reliability of all scenarios for each scale, as shown in table 1. The last row of the table contains the α values for the entire scale, measuring the inter-rater reliability for all answers. We observe that the ME scale has high inter-rater reliability while the 100-level scale is less reliable. Also, participants using the ME scale tend to exhibit higher agreement regarding the FP and FN cases and systematically disagree on the rejected cases. For the 100-level scale, we observe that participants have low agreement on all scenarios.

We analyze the validity of the ME scale by comparing the median normalized magnitude estimates with the median 100-level scores for each question set. Figure 1 presents the correlation plot between the two scales. A Shapiro-Wilk test indicates that the data of both scales do not follow a normal distribution ($p < 0.05$). Thus, we use the Spearman and Kendall rank correlation coefficients since these are non-parametric tests. Spearman returned a 0.98 and Kendall a 0.89 correlation between the ME and the 100-level scales ($p < 0.05$). Finally, a Mann-Whitney U test between the ME and 100-level scales gives a large p-value, indicating no statistically significant difference between the two scales.

4.2 Total Model Value due to Threshold

We evaluate the $V(\tau)$ function (i.e., the value at different rejection thresholds) using the values from the survey study obtained using the ME scale. We train three different binary hate speech classification models on the Waseem and Hovy [67] dataset. The used models are Logistic Regression (LR) with Character N-gram [67], a Convolutional Neural Network (CNN) based on Agrawal and Awekar [1], and a DistilBERT transformer [58]. We use Temperature Scaling to calibrate the CNN and the DistilBERT models following the approach from Guo et al. [33]. The model predictions are based on two different test datasets: the *seen* dataset and the *unseen* dataset. The *seen* dataset is the test set of Waseem and Hovy [67] and the *unseen* dataset is a test set from a separate but similar source [8]. We use the *unseen* dataset to simulate how the models would perform in a more challenging, realistic use case. Using unseen data that is similar but separate from the training set, we also investigate the impact of bias. Finally, we calculate the total value as a function of the threshold, $V(\tau)$, for all models with the reject option at all possible rejection thresholds (τ). When $\tau \in [0.0, 0.5]$, all predictions are accepted since the confidence of all predictions is above 0.5 in the case of binary classification. On the other hand, $\tau = 1.0$ implies that all predictions are rejected. We use the v values of the ME scale from table 1 to plot the results of all three models in figures 2a and 2b using equation (3). The diamond-shaped markers indicate the optimal confidence thresholds for rejection at which the model achieves the highest total value.

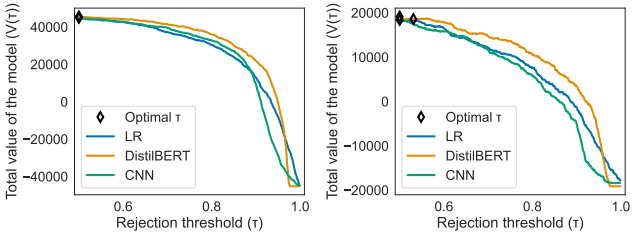
Participants ascribe higher absolute values to TP and TN scenarios compared to FP and FN ones (see table 1), which results in all but one model having the highest value when all predictions are accepted (see figures 2a and 2b). The rejection rates (i.e., the percentage of rejected predictions) and accuracies of accepted predictions at the optimal threshold across the three classifiers can be seen in the first two rows of table 2. If we were to take the view that the users' baseline expectation is correct machine decisions, then we can set the value of TP and TN to 0.0 and repeat our analysis to examine how $V(\tau)$ behaves as we consider only punishing incorrect predictions without rewarding correct predictions made

	LR			DistilBERT			CNN		
	τ	Acc	RR	τ	Acc	RR	τ	Acc	RR
Seen data	0.500	0.853	0.000	0.500	0.853	0.000	0.500	0.845	0.000
Unseen data	0.531	0.646	0.043	0.500	0.643	0.000	0.500	0.624	0.000
Seen data ($V_{TP}=0, V_{TN}=0$)	0.829	0.925	0.316	0.786	0.923	0.202	0.815	0.934	0.299
Unseen data ($V_{TP}=0, V_{TN}=0$)	0.999	0.818	0.991	0.974	1.000	0.996	0.961	0.833	0.980

Table 2: The optimal rejection thresholds (τ), the accuracy of the accepted predictions (Acc), and rejection rates (RR) of all models for both datasets using the values from the survey.

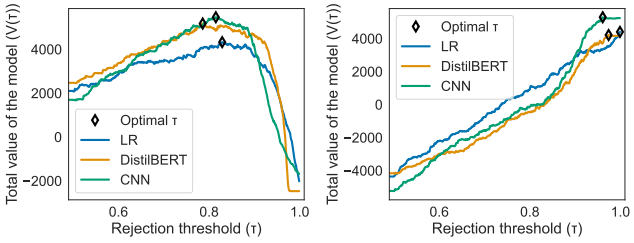
	LR		DistilBERT		CNN	
	$V(\tau_O)$	Acc	$V(\tau_O)$	Acc	$V(\tau_O)$	Acc
Seen data	45534	0.853	45250	0.853	44893	0.845
Unseen data	18563	0.631	19132	0.643	18385	0.624
Seen data ($V_{TP}=0, V_{TN}=0$)	4325	0.853	5172	0.853	5460	0.845
Unseen data ($V_{TP}=0, V_{TN}=0$)	4404	0.631	4213	0.643	5291	0.624

Table 3: The total values $V(\tau_O)$ and the accuracies (Acc) of all models. Here, τ_O is the optimal rejection threshold.



(a) evaluated on *seen* data **(b) evaluated on *unseen* data**

Figure 2: $V(\tau)$ curves of all models with v of TP=18.15, TN=36.32, FP=-16.69, FN=-28.08, and rejection=4.82.



(a) evaluated on *seen* data **(b) evaluated on *unseen* data**

Figure 3: $V(\tau)$ curves of all models with v of TP=0.0, TN=0.0, FP=-16.69, FN=-28.08, and rejection=4.82.

by the model (considering the regulation effect discussed in section 2). Figures 3a and 3b demonstrate that the optimal values are achieved at increased rejection thresholds (τ). The last two rows of table 2 show that the optimal τ values result in higher accuracies for the *seen* data while rejecting 31.6% of predictions. For the *unseen* data, we achieve high accuracies but reject a large fraction of the predictions.

We also compare the effect of using value and the widely-used accuracy metric in selecting the best model, shown in table 3. We observe that both metrics return the same optimal model when correct predictions are rewarded, though there is a difference between *seen* and *unseen* cases. When only incorrect predictions are punished, the optimal models are different as measured by the two metrics: in the case of *seen* data, both LR and DistilBERT perform better than CNN when measured by accuracy, while CNN delivers the highest value; the same observation holds true in the case of *unseen* data – where the optimal model switches from DistilBERT to CNN when we consider the value they deliver instead of accuracy.

5 DISCUSSION

5.1 Value Ratios, Reliability, and Validity

Our results show that TP and TN scenarios are highly valued. Participants seem to value correct predictions more than incorrect predictions across all scenarios, regardless of whether they are positive or negative. The value of rejected predictions is the closest to 0 (neutral), as expected, due to them not contributing any benefit or harm, but just delaying the publishing of the post due to the additional human moderation effort. For both scales, we observe the same relation of scenarios in terms of values (FN<FP<Rejection<TP<TN). The fact that correct decisions receive higher value ratings indicates strong user appreciation of correct machine decisions. The value of FN having a larger magnitude than the value of FP is noteworthy, as users appear to be more negatively affected when a non-hateful post is subject to moderation than when an instance of hate speech is classified as non-hateful. This implies that users would rather contend with an instance of hate speech than have an innocent user punished for a non-hateful post. This phenomenon may be explained by the Blackstone principle from the domain of criminal law: "Better that ten guilty persons escape, than that one innocent suffer" [20]. However, we do consider it surprising that the value of TN is greater than the value of TP. One possible reason could be

that people disagree more on what is considered hateful among the TP scenarios. We also encountered this phenomenon in the survey results where most people rated TN cases as non-hateful, while for the TP cases there were more disagreements.

Regarding reliability, Krippendorff's alpha, α, for the 100-level scale being lower than the one for the ME scale is unexpected, as the 100-level scale is bounded with fewer possible options. The stronger agreement for the ME scale indicates that it is indeed suitable for this task. Since α compares the expected difference with the observed difference, it follows that the alpha values for the entire scale should be greater than for the individual scenarios. Generally, participants tend to have low agreement on TP, TN, and rejection cases while they have a high agreement regarding the FP and FN cases. Users tend to agree more regarding what constitutes a misclassified instance than what constitutes a correctly classified instance. For the ME scale, we even observe systematic disagreement for the rejection case, as can be seen by its negative α value. This indicates that users are lower in agreement than one would expect by chance, showing the wide variety of opinions regarding rejection cases by users. By considering all answers, instead of answers for certain scenarios, we observe a greatly increased α, as the observed difference between ratings is closer to the difference expected by chance. For example, participants tend to agree on the classification of a single scenario, e.g. TP, but may give different values on both scales, resulting in lower α for the scenario but greater α across all scenarios. Beyond this, the low reliability for the positive compared to negative predictions indicates that participants disagree on what constitutes hate speech in the first place.

Regarding validity, we observe a strong correlation between scales, demonstrating that the ME scale is validated for measuring people's opinions about different hate speech detection scenarios. The almost S-shaped curve for the data points in figure 1 is due to the lower and upper bounds of the 100-level scale that restrict the participants' choices, making them more likely to assign the lowest or highest value. Meanwhile, the data points corresponding to the ME scale are skewed towards 0 because of the normalization.

5.2 Value Function for Rejection

The purpose of the reject option is to reject predictions where the risk of an incorrect prediction is too high. However, when we use all values obtained from the survey to measure the value function $V(\tau)$, the total value of a model with a reject option is maximized by accepting all predictions. As shown in figures 2a and 2b, values are positive at the beginning, decline steadily as the rejection threshold increases, and eventually become negative as more predictions are rejected. This observation is not surprising, as the absolute values of correct predictions are greater than the absolute values of incorrect predictions (see table 1).

However, instead of rewarding correct predictions, we believe it is more critical to emphasize penalizing incorrect predictions, as hate speech should be moderated effectively to minimize harm. To study the effects of this we also analyze the behavior of $V(\tau)$ when users do not experience an increase in value through correct classifications, i.e. TP and TN. To achieve this, we set the scenario values v of TP and TN equal to zero. This results in correct predictions effectively only increasing the total value by the v of rejection when accepted and decreasing when rejected, as can be seen in equation (3). The result in figure 3a shows a steady increase in value before it peaks for each of the three models, eventually falling again and becoming negative as almost all predictions are rejected. Hence, there is a strong incentive to reject some (but not all) predictions for the *seen* data. At the points where values are maximized, we found an optimal balance between accepting and rejecting predictions. Figure 3b shows that the values continually rise for all three models, only peaking as the rejection threshold approaches 1. This indicates that the model is very uncertain regarding its predictions for the *unseen* data, which may be expected. Initially, at the 0.5 rejection threshold, the value is negative as all predictions are accepted. When the rejection threshold increases, the value rises steadily since too many incorrect predictions are made. This indicates that the model is not performing well at the task (i.e., high confidence false predictions), and thus the optimal condition to reject most predictions makes the unviable model.

The results show that by penalizing incorrect predictions without rewarding correct predictions, a significant fraction of the predictions can be accepted from all three models. For unseen data, however, very few predictions from these models can be accepted and the majority are rejected. Such a result confirms the bias in the dataset as also found in previous studies [3, 32]. The results also show the utility of value as a metric in guiding the decision on when to reject machine predictions. Value utility is further confirmed in the results in table 3 from our experiment on optimal model selection: the best model selected by value is different compared to using accuracy as the metric.

5.3 Findings, Implications, and Limitations

Our survey study uncovers several interesting findings. First, social media users are more appreciative of correct decisions made by the platform, with an absolute magnitude higher than the (negative) perception of incorrect decisions. Among the correct decisions, users especially appreciate that non-hateful content is correctly identified and not banned. On the other hand, users show a much higher agreement on the negative value of incorrect decisions than correct ones, indicating a strong consensus over the harm (from both identifying hateful content to be non-hateful, and vice versa). These results indicate that while users appreciate correct decisions, minimizing incorrect decisions remains an important task for social media platforms. On the methodological side, we also believe our proposal of using ME for rating human perception can be particularly relevant for research that aims to tackle social science problems through quantitative approaches, like machine learning.

By integrating value as a parameter into the human-AI collaboration framework for rejecting machine decisions, we show that value can help guide the decision of when to accept machine decisions to reach the optimal value a model can deliver. By showing how the number of acceptable machine decisions changes when the model is applied to a dataset different from the training data, our results confirm findings from previous research that such datasets are biased and hence the trained models are as well. Our results also show that when considering value as an optimization target, the best model selected can be different compared with using accuracy as the metric. We believe these findings can benefit the research

community and industry alike, as they present a novel way of using a value-sensitive reject option to increase the utility of human-AI collaboration across domains.

Our work is limited to a relatively small sample size (68 subjects per scale). We expect the results to be more reliable at a larger sample size. Besides, optimal confidence threshold determination relies heavily on empirical data, which may not be available in real applications. An easier way for selecting the optimal threshold would be using well-calibrated models, for which the optimal threshold is only dependent on the human-perceived value. Although techniques such as Temperature Scaling can help improve the calibration of existing neural networks or transformer models such as DistilBERT, we still observe that all models are predisposed to producing high-confidence errors. Finally, due to taking the users' standpoint, we do not fully capture the cost of the moderation team being exposed to hate speech. We leave this as possible future work.

6 RELATED WORK

6.1 Hate Speech Detection

Online hate speech content refers to "online messages demeaning people on the basis of their race/ethnicity, gender, national origin, or sexual preference" [41]. Its characterizing features are properties of the target of the language, as compared to other types of online conflictual languages, which are defined by the intention of the author such as cyberbullying or flaming [11, 54]. A large body of discussion can be found on conflictual languages from social sciences, political science, and computer science [44, 63, 66]. Hate speech-related research in computer science has identified mismatches between the formalization of hateful content and how people perceive such languages [4]. These mismatches conceptually are further reflected in the technical biases of the machine learning systems used for filtering hateful content. For instance, Gröndahl et al. [32] found that F1 scores were reduced by up to 69% when training a hate speech detection model on one dataset and evaluating it using another dataset from a similar source. Similarly, Arango et al. [3] found that most research in hate speech detection overestimates the performance of the automated methods due to dataset bias. In response to these findings, our work aims to explore a human-AI collaborative approach for effective hate speech detection.

6.2 Human-AI Collaboration and Rejection

Human-AI collaboration aims to exploit the complementarity between the cognitive ability of humans and the scalability of machines to solve complex tasks at scale [6, 65]. Some work proposed new ways of collaboration, such as learning crowd vote aggregation models from features of the crowd task [36] and leveraging crowds to learn features of ML models [15, 56]. Recent work has shifted attention to human involvement in providing interpretations of model decisions and evaluating these interpretations [40, 55].A notable idea for hybrid human-AI decision-making was recently proposed by Callaghan et al. [12]: humans are involved after a machine decision is observed to have low confidence. Following works can be categorized in several dimensions, namely *when* rejection happens, on *what models*, and based on *what criteria* [34]. Regarding the "when", rejection can be implemented in three ways: the preemptive way where whether a data item needs to be handled by a human is decided beforehand [16]; the integrated way which uses a rejector inside the machine learning model (e.g., a rejection layer in a neural network) to decide whether a decision should be rejected [27]; and the dependent way, which is also the most common, which analyzes the rejection option after model decisions [18, 26, 31]. In terms of "what models", work has been done on rejecting decisions made by a range of models, such as SVMs [16, 31] and different neural networks [18, 27]. In our case, we apply the dependent way to reject models that are based on neural networks. In terms of "what criteria", Geifman and El-Yaniv [26] proposed a rejection function based on a predefined risk value, an idea also explored in [51]. But unlike ours, their proposals do not consider the impact of machine decisions in a specific context. The most relevant proposal to our work is from De Stefano et al. [18], who studied a confidence metric for determining the optimal rejection threshold. In their work, the threshold is calculated with simulations based on a set of predictions. Going beyond defining cost values from simulations, our approach determines cost values based on users' perception of machine decisions using a survey study with crowd workers.

6.3 Value Assessment and Measurement

Value is generally defined as desirable properties of an entity [9]. Specifically for machine learning systems Yurrita et al. [69] have identified relevant properties, including individual empowerment, conservation, universalism, and openness. Examples include outlining ethical principles of algorithmic systems [23], developing value-based assessment frameworks [69], and proposing new metrics for evaluating machine learning systems that incorporate value parameters [13]. However, a research gap in measuring value in social contexts has been identified by Olteanu et al. [52], who investigated human-centered metrics for machine learning evaluation in hate speech detection. Their work highlights the gap between accuracy-based evaluation metrics and user perception. Our work represents a first step towards filling the gap in the context of hate speech detection using ME with a crowdsourced survey.

7 CONCLUSIONS

This paper studies the operationalization and integration of value into human-AI collaboration for hate speech detection. We introduce a value-sensitive rejection mechanism for machine decisions that takes into account the implications of decisions from a user-centered standpoint. We propose ME to measure users' value perception regarding different hate speech detection scenarios. To validate ME, we design a survey study, showing that it can provide a reliable, human-centered assessment of the value a machine learning model delivers. Our survey study uncovers a series of interesting findings on user perception. In particular, participants appreciate correct decisions made by the platform, while they show a strong consensus over the harm of incorrect decisions. Our results show that value assessment performed by means of ME can guide us to select the best confidence threshold for rejecting machine decisions, thereby maximizing model value and potentially leading to a different best model than when using accuracy.

REFERENCES

[1] Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*. Springer, 141–153.

[2] I Elaine Allen and Christopher A Seaman. 2007. Likert scales and data analyses. *Quality progress* 40, 7 (2007), 64–65.

[3] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*. 45–54.

[4] Agathe Balayn, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. 2021. Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature. *ACM Transactions on Social Computing (TSC)* 4, 3 (2021), 1–56.

[5] Emilio Balda, Arash Behboodi, and Rudolf Mathar. 2020. Adversarial Examples in Deep Neural Networks: An Overview. In *Deep Learning: Algorithms and Applications*. 31–65.

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 13 (May 2021), 11405–11414. https://doi.org/10.1609/aaai.v35i13.17359

[7] Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72, 1 (1996), 32–68.

[8] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 54–63.

[9] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590* (2021).

[10] Harry N Boone and Deborah A Boone. 2012. Analyzing likert data. *Journal of extension* 50, 2 (2012), 1–5.

[11] Victoria K Burbank. 1994. Cross-cultural perspectives on aggression in women and girls: An introduction. *Sex Roles* 30, 3 (1994), 169–176.

[12] William Callaghan, Joslin Goh, Michael Mohareb, Andrew Lim, and Edith Law. 2018. MechanicalHeart: A Human-Machine Framework for the Classification of Phonocardiograms. In *CSCW'18*, Vol. 2. 28:1–28:17.

[13] Fabio Casati, Pierre-André Noël, and Jie Yang. 2021. On the Value of ML Models. *arXiv preprint arXiv:2112.06775* (2021).

[14] Justin Cheng and Michael S Bernstein. 2015. Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 600–611.

[15] Justin Cheng and Michael S. Bernstein. 2015. Flock: Hybrid Crowd-Machine Learning Classifiers. In *CSCW'15* (Vancouver, BC, Canada).

[16] Lize Coenen, Ahmed KA Abdullah, and Tias Guns. 2020. Probability of default estimation, with a reject option. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 439–448.

[17] Mary Cummings. 2006. Integrating ethics in design through the value-sensitive design approach. *Science and engineering ethics* 12 (11 2006), 701–15. https://doi.org/10.1007/s11948-006-0065-0

[18] Claudio De Stefano, Carlo Sansone, and Mario Vento. 2000. To reject or not to reject: that is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 30, 1 (2000), 84–94.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[20] Daniel Epps. 2014. The consequences of error in criminal justice. *Harv. L. Rev.* 128 (2014), 1065.

[21] EU. 2016. The EU Code of conduct on countering illegal hate speech online. *European Commission* (May 2016). https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en Visited on 07/03/2022.

[22] Karen Fitzner. 2007. Reliability and validity a quick review. *The Diabetes Educator* 33, 5 (2007), 775–780.

[23] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication* 2020-1 (2020).

[24] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–30.

[25] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. Unesco Publishing.

[26] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4885–4894.

[27] Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2151–2159. https://proceedings.mlr.press/v97/geifman19a.html

[28] Noah Giansiracusa. 2021. Facebook Uses Deceptive Math to Hide Its Hate Speech Problem. *Wired* (Oct 2021). https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/ Visited on 07/03/2022.

[29] Michael Gilliland. 2020. The value added by machine learning approaches in forecasting. *International Journal of Forecasting* 36, 1 (2020), 161–166. https://doi.org/10.1016/j.ijforecast.2019.04.016 M4 Competition.

[30] Michael Wojatzki Tobias Horsmann Darina Gold and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. (2018).

[31] Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, and Stéphane Canu. 2008. Support Vector Machines with a Reject Option. In *Advances in Neural Information Processing Systems*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), Vol. 21. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2008/file/3df1d4b96d8976ff5986393e8767f5b2-Paper.pdf

[32] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is "love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*. 2–12.

[33] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *ICML'17 - Volume 70*. 1321–1330.

[34] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. 2021. Machine Learning with a Reject Option: A survey. *arXiv preprint arXiv:2107.11277* (2021).

[35] Mathew Ingram. 2018. Facebook now linked to violence in the Philippines, Libya, Germany, Myanmar, and India. *Columbia Journalism Review* (Sep 2018). https://www.cjr.org/the_media_today/facebook-linked-to-violence.php "Visited on 07/03/2022".

[36] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining Human and Machine Intelligence in Large-scale Crowdsourcing. In *AAMAS'12 - Volume 1* (Valencia, Spain). 467–474.

[37] Shervin Khodabandeh, David Kiron, Françoise Candelon, Michael Chu, and Burt LaFountain. 2020. Expanding AI's Impact With Organizational Learning. MIT Sloan Management Review and Boston Consulting Group.

[38] Kate Klonick. 2018. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review* 131 (2018), 1598.

[39] Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research* 30, 3 (2004), 411–433.

[40] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684. https://doi.org/10.1177/2053951718756684 arXiv:https://doi.org/10.1177/2053951718756684

[41] Roselyn J Lee-Won, Tiffany N White, Hyunjin Song, Ji Young Lee, and Mikhail R Smith. 2020. Source magnification of cyberhate: Affective and cognitive effects of multiple-source hate messages on target group members. *Media Psychology* 23, 5 (2020), 603–624.

[42] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 1–32.

[43] Mujib Mashal, Suhasini Raj, and Hari Kumar. 2022. As Officials Look Away, Hate Speech in India Nears Dangerous Levels. *The New York Times* (Feb 2022). https://www.nytimes.com/2022/02/08/world/asia/india-hate-speech-muslims.html Visited on 07/03/2022.

[44] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.

[45] Mick McGee. 2004. Master usability scaling: magnitude estimation and master scaling applied to usability measurement. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 335–342.

[46] Howard R Moskowitz. 1977. Magnitude estimation: notes on what, how, when, and why to use it. *Journal of Food Quality* 1, 3 (1977), 195–227.

[47] Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Wilson Leão, Steeven Janny, and Christian Gagné. 2018. Attended Temperature Scaling: A Practical Approach for Calibrating Deep Neural Networks. https://doi.org/10.48550/ARXIV.1810.11586

[48] Paul Mozur. 2018. A Genocide Incited on Facebook, With Posts From Myanmar's Military. *The New York Times* (Oct 2018). https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html Visited on 07/03/2022.

[49] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. Survey research in HCI. *Ways of Knowing in HCI* (2014), 229–266.

[50] Karsten Müller and Carlo Schwarz. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association* 19, 4 (2021), 2131–2167.

[51] Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 2009. Accuracy-Rejection Curves (ARCs) for Comparing Classification Methods with a Reject Option. In *Proceedings of the third International Workshop on Machine Learning in Systems Biology (Proceedings of Machine Learning Research, Vol. 8)*, Sašo Džeroski, Pierre Guerts, and Juho Rousu (Eds.). PMLR, Ljubljana, Slovenia, 65–81. https://proceedings.mlr.press/v8/nadeem10a.html
[52] Alexandra Olteanu, Kartik Talamadupula, and Kush R. Varshney. 2017. The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection. In *Proceedings of the 2017 ACM on Web Science Conference* (Troy, New York, USA) *(WebSci '17)*. Association for Computing Machinery, New York, NY, USA, 405–406. https://doi.org/10.1145/3091478.3098871
[53] Maithra Raghu, Katy Blumer, Greg Corrado, Jon M. Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. *CoRR* abs/1903.12220 (2019). arXiv:1903.12220
[54] Charlotte Rayner and Helge Hoel. 1997. A summary review of literature relating to workplace bullying. *Journal of community & applied social psychology* 7, 3 (1997), 181–191.
[55] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[56] Carlos Rodriguez, Florian Daniel, and Fabio Casati. 2014. Crowd-Based Mining of Reusable Process Model Patterns. In *Business Process Management*. 51–66.
[57] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. 2018. On fine-grained relevance scales. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 675–684.
[58] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).
[59] Burcu Sayin, Fabio Casati, Andrea Passerini, Jie Yang, and Xinyue Chen. 2022. Rethinking and Recomputing the Value of ML Models. https://doi.org/10.48550/ARXIV.2209.15157
[60] Burcu Sayin, Jie Yang, Andrea Passerini, and Fabio Casati. 2021. The Science of Rejection: A Research Area for Human Computation. https://doi.org/10.48550/ARXIV.2111.06736
[61] Olivia Solon. 2017. Facebook is hiring moderators. But is the job too gruesome to handle? https://www.theguardian.com/technology/2017/may/04/facebook-content-moderators-ptsd-psychological-dangers, Last accessed on 2022-06-21.
[62] Stanley Smith Stevens. 1956. The direct estimation of sensory magnitudes: Loudness. *The American journal of psychology* 69, 1 (1956), 1–25.
[63] Alexander Tsesis. 2001. Hate in cyberspace: Regulating hate speech on the Internet. *San Diego L. Rev.* 38 (2001), 817.
[64] Heidi Tworek and Patrick Leerssen. 2019. An Analysis of Germany's NetzDG Law.
[65] Jennifer Wortman Vaughan. 2017. Making better use of the crowd: How crowdsourcing can advance machine learning research. *The Journal of Machine Learning Research* 18, 1 (2017), 7026–7071.
[66] Jeremy Waldron. 2012. The harm in hate speech. In *The Harm in Hate Speech*. Harvard University Press.
[67] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.
[68] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to Complement Humans. In *IJCAI'20*. 1526–1533.
[69] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *ACM Conference on Fairness, Accountability, and Transparency*.
[70] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 194 (nov 2018), 23 pages. https://doi.org/10.1145/3274463

A SURVEY

A.1 Variables

The independent variables are the possible scenarios (TP, TN, FP, FN, and rejection). We inform participants in the survey that when hate speech is detected, SocialNet ranks the hateful post lower so that it takes much more effort for the users to find the post. For the rejection scenario, we inform the participants in the survey that a moderator needs to check the post within 24 hours, and meanwhile, the post remains visible.The design decision of using 24 hours is based on the German NetzDG law, which allows the government to fine social media platforms if they do not remove illegal hate speech within 24 hours [64]. Our study has two control variables: the measurement scales and the content of posts. Regarding scales, as described before, we choose ME as our primary scale and use the 100-level scale for validation. Our dependent variables are reliability, validity, and value ratios. We use Krippendorff's alpha to compute reliability, where a value equal to or larger than 0.8 and 0.6 indicates reliable and tentative conclusions, respectively [39, 42]. Regarding validity, we use convergent validity [22] between the two scales to assess if they measure the same phenomenon. The value ratio variable describes the perceived value of the scenarios, which is measured by calculating the median of the normalized magnitude estimates of each decision scenario.

A.2 Demographics

We perform a demographic analysis regarding the sex, student status, continent of residence, nationality, language, and ethnicity of our participants to study whether there are statistical differences in how they perceive hate speech. We have multiple groups for nationality, ethnicity, and language and two groups for the features of student status, sex, and continent of residence (since only two continents were present in the demographic data of all participants).

Table 4 shows the group differences for the aggregated scenario types (TP, TN, FP, FN, and REJ) and highlights the p values that differ substantially. We found the most significant differences across all scenarios for different nationalities and languages of participants.

A.3 Pilot Study

We conducted a pilot study with 24 participants to test the survey and estimate the required completion time. The pilot study showed low inter-rater reliability, which may be due to unmatched expectations and unclear scenario descriptions. Initially, our scenarios mentioned that machine-detected hateful posts would be removed, which could be controversial. Thus, we changed scenarios to rank the hateful posts lower on the feed. Moreover, we added more lengthy descriptions to explain the potential consequences of all scenarios in a neutral tone, which prepares the participants to focus on evaluating harm (instead of giving rewards).

A.4 Example Tasks

WARNING: GRAPHIC LANGUAGE.

A.4.1 FN scenario with ME scale.

Suppose we have the following post:
"I want this stupid bitch out of my head I've never even liked her lmao I just basically want to rape her" (from Basile et al. [8])

Question 1:
Please indicate whether you feel that this post is hateful or not hateful.
Select input with options: 'hateful' or 'not hateful'

Question 2:
You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet's automated detection system is confident that it is not hateful.

- This post remains visible on other people's main feeds as well.

	Two groups			More than two groups		
	Sex	**Student**	**Continent**	**Nationality**	**Language**	**Ethnicity**
TP	0.302	**0.032**	0.286	0.218	0.109	0.242
TN	0.726	0.379	0.204	0.190	0.216	0.281
FP	0.699	0.933	0.073	**0.020**	**0.040**	**0.037**
FN	0.961	0.150	0.847	0.478	0.438	0.584
REJ	0.835	0.625	0.496	0.271	0.103	0.068

Table 4: Statistical differences between different groups of participants for various demographic characteristics for each aggregated scenario type in the ME survey. Each cell contains the p value of either the Mann-Whitney U test for two groups or the Kruskal-Wallis test for more than two groups. Bold text indicates significant statistical differences.

- If the post is hateful after all, then the detection system was incorrect. This hateful post is now visible on people's main feeds with the consequence that they can get harmed.
- If the post is indeed not hateful, then the detection system was correct.
- You should be aware that it is expected for SocialNet to correctly classify hate speech. Wrong classifications are undesirable as they may cause harm to people.

Please indicate whether you agree, disagree, or are neutral about SocialNet's decision.
Select input with options: 'agree', 'neutral', or 'disagree'.

Question 3 (only visible when 'agree' or 'disagree' are selected):
Please indicate how much you agree/disagree with SocialNet's decision using any positive number.
Show ME input field that allows all positive values greater than 0.

A.4.2 FP scenario with 100-level scale.

Suppose we have the following post: *"Ontario concerns were not listened at immigration meeting in Ottawa For illegal crossing"* (from Basile et al. [8])

Question 1:
Please indicate whether you feel that this post is hateful or not hateful.
Select input with options: 'hateful' or 'not hateful'

Question 2:
You are a user of the SocialNet platform and have not seen this post on your main feed because SocialNet's automated detection system is confident that it is hateful.

- You can still find this post when you scroll down your feed since SocialNet ranks hateful posts lower.
- If the post is not hateful after all, then the detection system was incorrect. This neutral post is now ranked lower on people's feeds with the consequence that the post cannot easily reach the author's followers.
- If the post is indeed hateful, then the detection system was correct.
- You should be aware that it is expected for SocialNet to correctly classify hate speech. Wrong classifications are undesirable as they may cause harm to people.

Please indicate whether you agree, disagree, or are neutral about SocialNet's decision. **Select input with options: 'agree', 'neutral', or 'disagree'.**

Question 3 (only visible when 'agree' or 'disagree' are selected):
Please indicate how much you agree/disagree with SocialNet's decision using any positive number from 1 to 100. If you feel neutral about SocialNet's decision, select neutral in the field above.
Show a numerical slider with values between 1 and 100.

A.4.3 Rejection scenario with 100-level scale.

Suppose we have the following post: *"Ever been so hungover that your stomach feels like it's eating itself"* (from Basile et al. [8])

Question 1:
Please indicate whether you feel that this post is hateful or not hateful.
Select input with options: 'hateful' or 'not hateful'

Question 2:
You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet's automated detection system was not confident enough in whether it was hateful or not.

- An internal human moderator at SocialNet needs to look at it within at most 24 hours.
- Meanwhile, the post remains visible on people's main feeds.

Please indicate whether you agree, disagree, or are neutral about SocialNet's decision.
Select input with options: 'agree', 'neutral', or 'disagree'.

Question 3 (only visible when 'agree' or 'disagree' are selected):
Please indicate how much you agree/disagree with SocialNet's decision using any positive number.
Show a numerical slider with values between 1 and 100.