Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity

Samuel Carton, Qiaozhu Mei, Paul Resnick
University of Michigan, Ann Arbor, Michigan
{scarton, qmei, presnick}@umich.edu

Abstract

We present an experimental assessment of the impact of feature attribution-style explanations on human performance in predicting the consensus toxicity of social media posts with advice from an unreliable machine learning model. By doing so we add to a small but growing body of literature inspecting the utility of interpretable machine learning in terms of human outcomes. We also evaluate interpretable machine learning for the first time in the important domain of online toxicity, where fully-automated methods have faced criticism as being inadequate as a measure of toxic behavior.

We find that, contrary to expectations, explanations have no significant impact on accuracy or agreement with model predictions, through they do change the distribution of subject error somewhat while reducing the cognitive burden of the task for subjects. Our results contribute to the recognition of an intriguing expectation gap in the field of interpretable machine learning between the general excitement the field has engendered and the ambiguous results of recent experimental work, including this study.

Introduction

Interpretable machine learning seeks to explain the predictions of machine learning models in human-understandable terms. There are a number of reasons why this quality might be desired, but many of them reduce to concerns about robustness–that even very powerful modern techniques (i.e. deep neural nets) are not reliable enough to be applied with complete autonomy to high-stakes decisions. The hope is that increasing the transparency of these models will allow human auditors to recognize when they are liable to make mistakes (Doshi-Velez and Kim 2017).

However, the contemporary interpretable machine learning literature has generally been characterized by a profusion of technique and a paucity of evaluation. While many methods have been proposed in recent years, they have tended to be accompanied by ad-hoc evaluations consisting primarily of proxy empirical metrics and case studies (e.g. deepLIFT (Shrikumar, Greenside, and Kundaje 2017), integrated gradients (Sundararajan, Taly, and Yan 2017)). User studies, when present, tend to be limited in scale and focused on proxy metrics (e.g. LIME (Ribeiro, Singh, and Guestrin 2016), MUSE (Lakkaraju et al. 2019)).

So while many techniques have been proposed, it remains unclear how interpretability can actually improve the way humans use model advice to make decisions.

The task of detecting toxic online behavior is a good example of a domain where these robustness concerns have become very salient. Perhaps the most ambitious application of machine learning to this problem to date has been the Perspective API[1], a public-facing toxicity classification model which was released in conjunction with a large dataset of crowd-labeled Wikipedia revision comments (Wulczyn, Thain, and Dixon 2017). While praised for its scope, the project has attracted criticism for its vulnerability to adversarial examples and poor generalization across platforms (Hosseini et al. 2017), as well as its systemic bias (Sap et al. 2019; Dixon et al. 2018). These specific criticisms have been accompanied by a general skepticism about whether automated classifiers can ever fully replace human oversight in content moderation (Blackwell et al. 2018b).

The need for robust performance on this challenging and important task and the perception of a need for some level of human oversight suggests that it is a good application area for interpretable machine learning. Trained models have one set of traits: speed; consistency; overall accuracy. Human annotators have another: the ability to understand nuance, context and edge cases. It is possible that interpretability might enable efficient hybridization of human and model effort in ways that outperform either agent alone (Bansal et al. 2019). Simultaneously, experimentation within this domain has the potential to improve the field's understanding of the human factors of interpretability and begin solving the problem of how to usefully apply machine learning explanations.

In this paper we evaluate the effectiveness of feature attribution-style explanations in helping humans make decisions about the toxicity of social media posts. We ask subjects to predict the *consensus toxicity* of such posts as a means of establishing an external standard of correctness for what otherwise is a subjective decision task.

[1] https://www.perspectiveapi.com

We sample comments from the Wikipedia talk page toxicity dataset of (Wulczyn, Thain, and Dixon 2017), representing a range of true toxicity scores and decision difficulties. We then conduct a 2×2 between-subject experiment that assesses the impact of adding 1) a model prediction as a visual element alongside the comment text, and 2) explanations for the predictions of that model via highlighting relevant words and phrases within the text. We also include two extension conditions that test variant explanation techniques; a "partial" variant that highlights a minimal amount of relevant text, and a "keyword" variant that only identifies toxic words without regard for context or phrase structure.

We investigate three research questions:

RQ1: Presence of model predictions. How does the advice of an unreliable predictive model affect subject performance in predicting consensus toxicity of social media comments?

RQ2: Presence of explanations. Do (attribution-style) explanations help users make better use of advice from such an unreliable model?

RQ3: Explanation type. Do more minimal "partial" or sparser "keyword" explanations exhibit different performance properties from explanations optimized for completeness?

In summary, the contributions of this paper are as follows:

- We test the feasibility of interpretable machine learning for semi-automated consensus toxicity detection.
- We add to a small but growing body of evidence suggesting that the most popular types of explanations aren't adequate to improving human performance on decision tasks.
- We test the relative effectiveness of three different approaches for extractive, feature-based explanations of text classifier decisions.

Related Work

Online Abuse

Abusive online language has recently attracted increased attention from the research community, including several recently-established annual workshops (Waseem et al. 2017; Kumar et al. 2018). It goes by many names and subcategories in the literature, including hate speech (Fortuna and Nunes 2018), aggression (Kumar et al. 2018), toxicity (Wulczyn, Thain, and Dixon 2017), cyberbullying (Hosseinmardi et al. 2015), harassment (Golbeck et al. 2017) and incivility (Anderson et al. 2016).

A number of large-scale datasets have been published in recent years capturing variants of these phenomena (Wulczyn, Thain, and Dixon 2017; Napoles et al. 2017; Golbeck et al. 2017; de Gibert et al. 2018; Hosseinmardi et al. 2015), while many papers have been published on the application of machine learning to the detection thereof (Salminen et al. 2020; Pavlopoulos, Malakasiotis, and Androutsopoulos 2017; Chancellor et al. 2017; Chandrasekharan et al. 2019).

However, the task of detecting abuse has nuances that prevent totally automatic methods from being a good solution on their own. The task is subjective and context-specific. Different communities have different norms for acceptable content (Chandrasekharan et al. 2018; Fiesler et al. 2018).

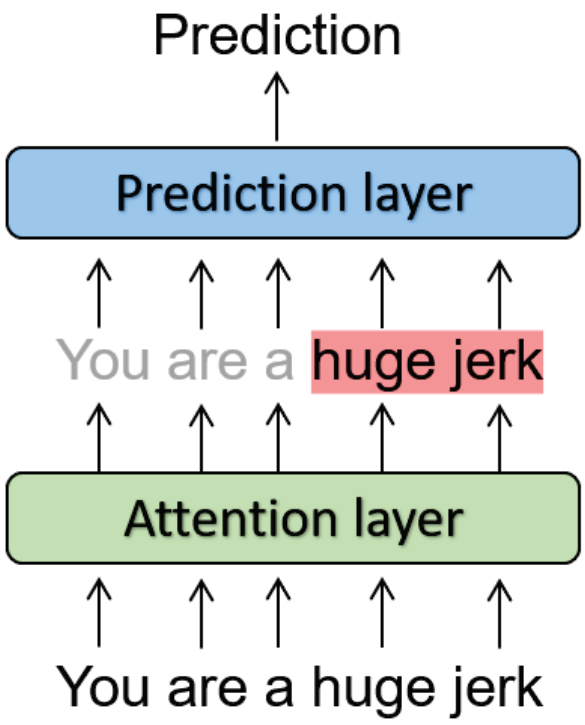

Figure 1: Explanatory model architecture.

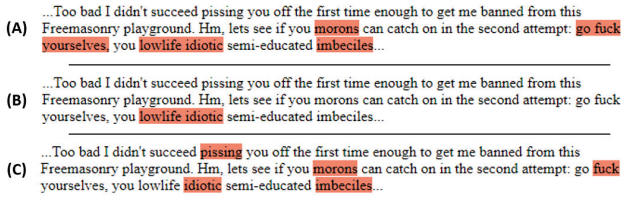

Figure 2: Example of explanation variants: (A) Full explanation; (B) Partial explanation; (C) Keyword explanation.

Different individuals have different perceptions about what constitutes abuse with respect to linguistic features like profanity (Malmasi and Zampieri 2018) or context (Blackwell et al. 2018a). It is very easy for labeller bias to propagate into trained models (Binns et al. 2017), while (Olteanu, Talamadupula, and Varshney 2017) point out that traditional metrics like accuracy may belie the actual human impact of model errors. Standard classifiers are also easy to fool with nonstandard language (Hosseini et al. 2017).

Interpretable Machine Learning

Interpretable machine learning seeks to extract insights about why models make their predictions. The most popular type of technique is feature attribution, which explains classifier decisions by noting which features of an individual item had what impact on the classifier's decision for that item (Murdoch et al. 2019).

This type of technique generally comes in three flavors: 1) perturbation-based methods like LIME (Ribeiro, Singh, and Guestrin 2016) or SHAP (Lundberg and Lee 2017); 2) gradient-based methods such as deepLIFT (Shrikumar, Greenside, and Kundaje 2017) and integrated gradients (Sundararajan, Taly, and Yan 2017); and 3) attention methods which explicitly model feature importances rather than calculating them retroactively (e.g. (Lei, Barzilay, and Jaakkola 2016)).

No universally accepted benchmark yet exists for attribution mask quality, though (DeYoung et al. 2019) is a prominent recent movement in this direction, focusing on the ideas of *comprehensiveness* and *sufficiency* as initially proposed by (Yu et al. 2019). Recurrent neural networks have been noted as producing incoherent results when subjected to perturbation-style analysis (Feng et al. 2018), while certain types of attention mechanism have been subject to a protracted debate over their utility and informativeness (Jain and Wallace 2019; Wiegreffe and Pinter 2019).

A few works have specifically pursued the idea of interpretable ML for abuse detection: (Svec et al. 2018) shows that an interpretable model can match human-generated annotations with high precision, while (Pavlopoulos, Malakasiotis, and Androutsopoulos 2017) proposes using explanations to help humans make decisions about borderline instances. (Wang 2018) analyzes pitfalls associated with using interpretable ML for abuse detection. Finally, (Carton, Mei, and Resnick 2018) gets both good precision and recall relative to human annotations by using an adversarial objective for producing attention masks.

Human Experimentation in Interpretability

There has been relatively little empirical work on how explanations affect human outcomes. (Abdul et al. 2018) and (Gillies et al. 2016) suggest that robust human experimentation can bridge this gap, while (Dove et al. 2017) calls generally for work that explores "the interplay between ML statistical intelligence and human common sense intelligence". (Doshi-Velez and Kim 2017) and (Gilpin et al. 2018) both lay out taxonomies of evaluation types for interpretability, calling for increased rigor in human evaluations particularly.

A small body of work has begun to emerge focused on experimentally evaluating interpretability techniques in terms of human outcomes. Examples of this style of work include (Lai and Tan 2019; Lage et al. 2019; Poursabzi-Sangdeh et al. 2018; Friedler et al. 2019; Weerts, van Ipenburg, and Pechenizkiy 2019). A commonality of these papers is that they generally demonstrate no significant effect on human performance from the presence of explanations.

Beyond interpretability, this study is an example of AI-advised human decision making, which has been shown to be a difficult and delicate partnership to enable (Bansal et al. 2019). More generally it falls into a genre of literature which might be termed human-AI interaction, which has shown recently that intelligent-yet-opaque algorithms tend to inspire both discomfort and inordinate trust (Springer, Hollis, and Whittaker 2017). This discomfort, at least, can be partially alleviated by increasing transparency (Eslami et al. 2018). (Jhaver et al. 2019) examines collaboration between moderators and syntax-based "automod" features on Reddit, noting a need for transparency in these tools.

Interpretable Classification Model

In order to generate toxicity predictions with feature attribution, we use the neural attention model described by (Carton, Mei, and Resnick 2018) (Figure 1). This is a hard-attention model which uses one recurrent neural net (RNN) to produce a discrete attention mask over the input text, and another RNN to make a prediction from the attention-masked text, with the two layers trained together to optimize a combination of predictive accuracy and mask sparsity. The model uses an adversarial mechanism that encourages it to include *all* toxic content in the attention mask so that, as far as it can distinguish, whatever is left behind rates as non-toxic, similar to the mechanism independently proposed by (Yu et al. 2019). The model predicts a target value for the whole text between 0 (nontoxic) and 1 (toxic).

An important quality of this model is that it prefers empty attribution masks when it believes the toxicity score of a comment is low. What this means is that low-predicted-toxicity comments are liable to have no highlighted content, leading to a semantic difference in how positive and negative predictions are explained by this model: positive predictions are justified with what the model considered toxic, while negative predictions are "justified" by a lack of evidence to the contrary.

As discussed above, there are a number of existing approaches for feature attribution. The (Carton, Mei, and Resnick 2018) model is a reasonable choice for this study because it was shown to perform well relative to alternatives on the task of identifying all personal attacks in social media text, and appears to avoid pitfalls that have been identified with other attribution approaches for recurrent neural networks (Feng et al. 2018; Jain and Wallace 2019).

We supplement the "full" attribution mechanism described above (Figure 2A) with two variants. In the first variant, we produce "partial" attribution masks by taking any multi-chunk mask produced by the model and reducing it to just the single chunk which maximizes the accuracy of the predictor when considering only that chunk. So when a comment has multiple discrete instances of toxicity, we reduce the explanation to just the most toxic instance (Figure 2B). This variant is intended to test the hypothesis that an explanation needs to consist only of *sufficient* information, not *comprehensive* information.

The second variant produces keyword-based explanations. We train a bag-of-words logistic regression classifier on the same dataset as the full model, and use the coefficients of this model to designate certain words (e.g. "pissing", "morons" in Figure 2C). This amounts to a dictionary approach, where certain words are always considered toxic and others always nontoxic. It produces very sparse explanations, where only the most toxic single words are highlighted, without regard for context or phrase structure. It also produces explanations that are not necessarily aligned with the model's predictions. This variant is intended to gauge the value of the relatively sophisticated RNN-based full attribu-

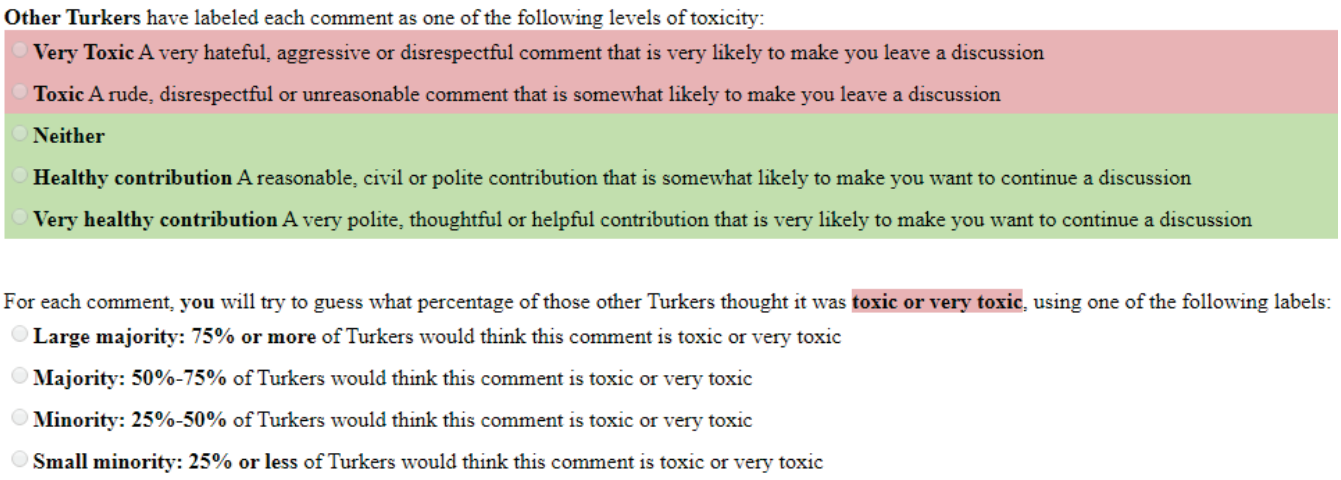

Figure 3: Instructions given to Phase 2 subjects, which also summarize Phase 1 task.

tion mechanism against a much simpler approach in terms of real human outcomes.

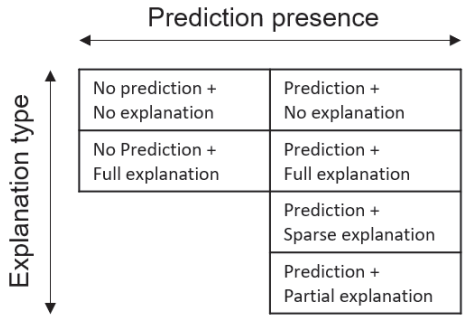

Figure 4: Phase 2 experimental conditions.

Experimental Design

The experiment sought to evaluate how well subjects predict the consensus toxicity of social media comments with varying levels of algorithmic assistance. It consisted of the following basic structure:

1. **Sample comments:** Draw a sample of comments from the (Wulczyn, Thain, and Dixon 2017) dataset, selecting for diversity in existing toxicity scores and model error. 96 comments sampled, split into 2 sets of 48 each.
2. **Collect consensus (Phase 1):** Collect consensus toxicity score for each comment by asking workers for their personal opinion of each comment. 54 subjects reviewed each of the 2 comment sets; 108 workers total.
3. **Predict consensus (Phase 2):** Ask subjects to predict outcome of Phase 1 with a varying level of algorithmic assistance. 40 subjects reviewed each of the 2 comment sets across 6 treatment conditions; 80 subjects total per condition; 480 subjects total.

The structure of the Phase 2 experiment was a 2×2 between-subject design (Figure 4) with two treatments: presence of prediction and presence of explanation, as well as two extension conditions in which the prediction is present with a variant explanation type, "keyword" and "partial" .

Subjects Subjects were recruited using the Amazon Mechanical Turk platform in August 2018. Subjects had to be US-based, and have completed at least 1000 HITs with 95% acceptance or more in order to qualify for the experiment. 588 total subjects participated, as enumerated above.

This subject count was chosen through a simulated power analysis to have a high (80%) chance of detecting an effect size of 0.05 in the primary outcome, accuracy, given outcome variances observed in a pilot study. This minimum detectable effect size was chosen as representing a 10% improvement on what was observed to be the baseline human accuracy of $\sim$50% on the task.

This study was approved by the University of Michigan institutional review board.

Comment Sampling

Each subject reviewed a sample of 48 comments drawn from the (Wulczyn, Thain, and Dixon 2017) dataset. This dataset consists of roughly 100,000 Wikipedia revision comments each labeled on a 5-point toxicity scale (Figure 3) by at least 10 workers on the CrowdFlower (now Figure Eight) platform. We followed (Wulczyn, Thain, and Dixon 2017) in binarizing each 5-point label to toxic(1)/nontoxic(0), and took the fraction of users who found the comment toxic to form a continuous toxicity score. Hence, a comment which 3/10 workers deemed toxic is assigned a 0.3 true toxicity label for the purpose of model training and evaluation.

In our study we convert this toxicity prediction task into a four-class classification task, with each class representing one quartile of the true toxicity score: large majority (75%

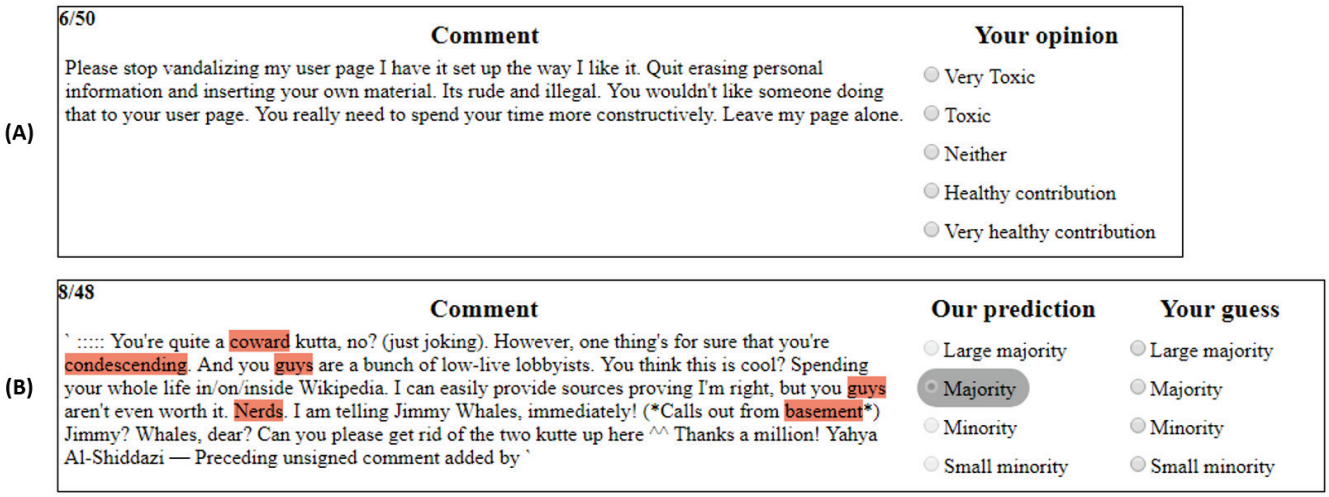

Figure 5: (A) Example comment in the Phase 1 personal opinion task; (B) Example comment in the Phase 2 prediction experiment

to 100%); majority (50% to 75%); minority (25% to 50%); and small minority (0% to 25%). We chose to frame the task this way rather than as a binary classification task in order to make it more difficult for human participants, and therefore to provide more room for improvement in accuracy.

Without resampling, this dataset is quite unbalanced. Roughly 90% of instances have a toxicity score below 0.5. Furthermore, it represents a relatively "easy" classification task: our LSTM classifier achieves 86% four-class accuracy (96% binary accuracy). We were interested in understanding human performance across the full range of true labels. Furthermore, we wanted to investigate whether explanations could allow human users to overturn classifier errors. Hence, in choosing which comments to present to our human subjects, it was necessary to use stratified sampling to oversample both toxic instances and classifier errors.

Specifically, we sampled 48 comments total for each comment set, split evenly across the 4 toxicity quartiles described above. For each quartile, we sampled 12 comments: 6 where our model predicted the correct quartile, and 2 each of where the model predicted each of the 3 other quartiles. For the two edge quartiles, large majority and small minority, there were not enough cases where the model predicted the other extreme. For these, we instead sampled 3 from the next most extreme error and only 1 from the most extreme.

Put together, this process resulted in a sample which is 50% toxic/nontoxic, 25% in each quartile, and on which our model achieves 50% classification accuracy at the quartile level (with respect to the labels present in the (Wulczyn, Thain, and Dixon 2017) dataset). Thus, subjects were presented with a roughly even number of comments that were toxic versus nontoxic, and a roughly even number for which the classifier was correct versus incorrect.

As a result of this process, we presented participants with a sample of comments on which the model is quite inaccurate. Our rationale for this design choice is our proposition that **for interpretability to be useful, it has to allow humans to overturn model mistakes more often than they disregard model successes**. In order to potentially show this effect, we needed a substantial number of model mistakes for subjects to recognize.

This is in contrast to similar studies such as (Poursabzi-Sangdeh et al. 2018) and (Lai and Tan 2019), where the model was more accurate than the human subjects. In those studies, a big improvement to human accuracy was possible simply by persuading subjects to agree with the model, so an effective explanation was one which increased user trust in the model output, an effect not necessarily due to real utility. In our study, no such avenue existed, closing off one threat to the internal validity of the study.

Phase 1: Ground-Truth Consensus (Re)Collection

In Phase 1, we recollected ground-truth consensus toxicity scores despite having access to an existing ground truth in the (Wulczyn, Thain, and Dixon 2017) dataset. We did so by having 54 subjects label each comment using the same questionnaire as (Wulczyn, Thain, and Dixon 2017), which asks the worker to rate the comment on a 5-point scale between "Very toxic" and "Very healthy" (Figure 5A). When we aggregated the results of this phase, we binarized each response into either toxic ("toxic" or "very toxic") or nontoxic (any other option), took the mean across subjects, and then bucketed each mean into the appropriate quartile to serve as the true four-class toxicity label for that comment.

We recollected these labels for several reasons. First, having 54 subjects for each comment instead of 10 meant a generally lower-variance true label for each comment. Second, drawing our ground truth from the same population as the Phase 2 subjects was more fair to them, since that phase involved asking subjects to make predictions about their own population rather than that of CrowdFlower.

The third reason is that because we sampled a disproportionate fraction of items where the classifier was incorrect, we were worried that a disproportionate number of these

	Condition	Accuracy		Agreement		False negative rate		False positive rate		Seconds/ comment	
		Mean	p	Mean	p	Mean	p	Mean	p	Mean	p
	Model	0.375		1		0.396		0.229			
1	No pred. + no exp.	0.544		0.432		0.276		0.179		10.16	
2	No pred. + full exp.	0.514	0.294^{1}	0.436	0.959^{1}	**0.353**	0.004^{1**}	**0.133**	0.034^{1*}	9.75	0.577
3	Pred. + no exp.	0.525	0.513^{1}	**0.535**	0.000^{1***}	0.315	0.154^{1}	0.16	0.389^{1}	**11.87**	0.045^{1*}
4	Pred. + full exp.	0.524	0.959^{3}	0.533	0.959^{3}	0.337	0.389^{3}	0.139	0.294^{3}	**9.95**	0.032^{3*}
5	Pred. + partial exp.	0.526	0.959^{4}	0.519	0.665^{4}	0.357	0.434^{4}	0.116	0.289^{4}	10.65	0.374^{3}
6	Pred. + keyword exp.	0.518	0.959^{4}	0.531	0.959^{4}	0.346	0.924^{4}	0.135	0.959^{4}	9.88	0.920^{4}

Table 1: Mean subject performance metrics across conditions. 80 subjects per condition. p-value superscripts indicate comparison condition. *:$p < 0.05$; **:$p < 0.01$; ***:$p < 0.001$.

items would be ones where the label itself was noisy due to random labeler error. The Phase 1 task, therefore, served to reduced this chance by re-surveying the toxicity of the sampled comments.

We chose to follow the (Wulczyn, Thain, and Dixon 2017) questionairre and define ground truth toxicity scores as a mean of binary responses for purposes of synchronicity with the model. If we recollected a ground truth generated differently from this dataset (and therefore drawn from a different distribution), the model's predictions and explanations would be tuned to a different data distribution than this ground truth, and this disjunction would represent a threat to the internal validity of the study.

Phase 1 Quality Assurance and Compensation Quality assurance for Phase 1 was via two attention checks in each question set. Subjects were made aware of the presence of the attention checks, though not of how many there were. Each attention check consisted of a sentence embedded within a comment asking the user to assign it a certain label chosen to be the opposite of the true label for that item. Workers thus were likely to miss the attention checks if they were putting random labels or failing to carefully read the comment texts.

Phase 1 workers were compensated with a base payment of $1.50 plus a bonus of $0.50 for each attention check they marked correctly. We discarded the results of any subject who missed both attention checks (3 in total).

Phase 2: Prediction Experiment

In Phase 2, we asked subjects to predict the outcome of Phase 1. Hence, if a comment was designated toxic by 60% of the subjects who reviewed it in Phase 1, the target class for that comment would have been "majority" in Phase 2.

The purpose of Phase 2 was to examine how well subjects were able to integrate advice from an error-prone model into their own predictions, and the extent to which explanations made them more or less effective in doing do.

As described above, each Phase 2 subject made predictions under one of six different experimental conditions (Figure 4): 1) No prediction + no explanation; 2) Prediction + no explanation; 3) No prediction + full explanation; 4) Prediction + full explanation; 5) Prediction + partial explanation; and 6) Prediction + keyword explanation.

Subjects were asked to review each text and choose one of the toxicity quartiles described above (Figure 5B).

In the control condition, workers made toxicity predictions without any algorithmic assistance. Two treatments were explored: the presence of the algorithmic predictions, and the presence of explanations in the form of word highlighting As described above, explanations came in three variants: full, partial and keyword-based (Figure 2).

In prediction-present conditions, the algorithm's prediction was presented to the right of the comment text (Figure 5B). In order to prevent workers from simply mirroring the model prediction, the instructions explained that the model was "not entirely reliable", and that workers would have to decide how much they wanted to rely on it. The phrase "your guess" was used to verbally distinguish model activity ("prediction") from human activity on the task.

In explanation-present conditions, the explanation was presented as red highlighting over the comment text (Figure 5B)). This feature was explained to users as the algorithm attempting to highlight toxic content.

Phase 2 Quality Assurance and Compensation Workers in Phase 2 were given a base payment of $1.25 plus a bonus of $0.05 for each item they predicted correctly relative to the aggregated results of Phase 1. We didn't use any other quality assurance mechanism for two reasons. First, we were relying on the natural desire of our subjects to maximize their earnings under the stipulation of the error-prone model. Second, because we were interested in measuring speed, we wanted to simulate a smoother perceived tradeoff between effort and reward. If we had included attention checks on this task, subjects would have been strongly incentivised to carefully read every token, which would have potentially masked any effect on subject speed arising from the presence of explanations.

Outcome Variables

We measured the accuracy and speed with which Phase 2 subjects made predictions about comment toxicity, as well as the extent to which they agreed with the model, which we treat as a behavioral indicator of model trust.

Results

Table 1 summarizes the results of the study, showing how accuracy, speed and agreement with the classifier varied across conditions.

Statistical Testing

We calculate effect sizes of each condition with respect to our three research questions. To assess the impact of predictions and explanations alone (RQ1), we compare "No prediction + full explanation" and "Prediction + no explanation" against "No prediction + no explanation". To assess the impact of explanations given the presence of a prediction (RQ2) we compare "Prediction + full explanation" against "Prediction + no explanation". Finally, to understand the relative impact of the two explanation variants (RQ3), we compare both "Prediction + partial explanation" and "Prediction + sparse explanation" against "Prediction + full explanation".

For every comparison we perform a two-tailed t-test. We report the p-value for each comparison, adjusted by Benjamini-Hochberg correction across the 5 comparisons and 5 outcomes with a target false discovery rate of 0.05.

Accuracy and Agreement

We find that the presence of the model's prediction has a marginal negative effect on the accuracy of subjects, an effect that does not vary significantly with the presence of explanations of any variant (Figure 6). The rightmost columns demonstrate that this result is being driven largely by users tending to believe the model over their own judgment. When the model is correct, human accuracy rises. When it is incorrect, human accuracy falls.

Figure 7 further shows this effect. Subjects tended to agree with the model prediction when it was visible, with explanations again making no significant difference. Breaking this result out across comments for which the predicted label was toxic or nontoxic, we find that explanations were liable to reduce user agreement with toxic predictions while improving user agreement with nontoxic predictions (which would have had no highlighting). Thus, while users were somewhat inclined to critique positive evidence, they were less inclined to question a lack of evidence.

False Positive Rate and False Negative Rate

The difference in how subjects perceived positive and negative evidence is mirrored somewhat in their false positive and false negative rates. While we find no significant effect of explanations on accuracy per se (Figure 6), breaking subjects errors down into false negatives and false positives shows they do impact the distribution of errors made by humans.

In particular, we find that explanations alone increase false negative rates while decreasing false positive rates relative to the completely unassisted condition (Figures 8 and 9). This result implies that feature attribution changes the way that subjects read the comments, making it easier for them to avoid errors of attribution but more liable to make errors of omission.

Speed

Figure 10 shows the speed effect of the various conditions. We find that the addition of a prediction adds a significant time penalty in comparison to the unassisted condition, presumably as users are forced to attempt to reconcile their own opinions with that of the classifier. However, adding explanations erases this time penalty, bringing the mean comment labeling time back down to that of the unassisted condition.

Outlier Analysis

Due to the lack of a hard quality assurance measure in Phase 2, the study had some vulnerability to unreliable subjects. To assess the potential impact of this, we try removing any subjects whose prediction accuracy was more than two standard deviations below the mean for their condition. Doing so removes a total of 22 subjects out of 480 (4.5%); 4 from condition 1; 5 from condition 2; 4 from condition 3; 4 from condition 4; 2 from condition 5; and 3 from condition 6.

Repeating the analysis above with the removed outliers produces very similar results. The only significant difference is that the observed time penalty associated with the presence of predictions is reduced slightly ($p = 0.045$ to $p = 0.067$). The relative speed improvement from explanations remains, however ($p = 0.041$).

Discussion

The results described above provide answers to our three research questions:

RQ1: Presence of Model Predictions

We find that the presence of a visible model prediction tends to bias subjects in favor of the prediction, whether it is correct or incorrect. There is also a significant speed penalty associated with the presence of a model prediction, as users are forced to ingest and reconcile an additional piece of information beyond the text itself.

This suggests that it is difficult for users to effectively integrate advice from a model into their decision-making. Insofar as subjects are liable to reject model advice, they are as likely to discard good advice as to reject bad advice in this study.

RQ2: Presence of Explanations

We find no significant effect of explanations on user accuracy or agreement when exposed to a model prediction. One possible explanation for this failure is that, in this domain, the difficulty in prediction lies not in identifying what words and phrases may be toxic, but in predicting exactly how those words and phrases are liable to be perceived by the general population.

Dividing subject errors into false negatives and false positives sheds more light on the situation. The model has both a high false negative rate and false positive rate compared to unassisted subjects, but both prediction and explanations alone raise the false negative rate and *lower* the false positive rate among human subjects. Explanations alone lower the FPR by a greater amount than predictions alone.

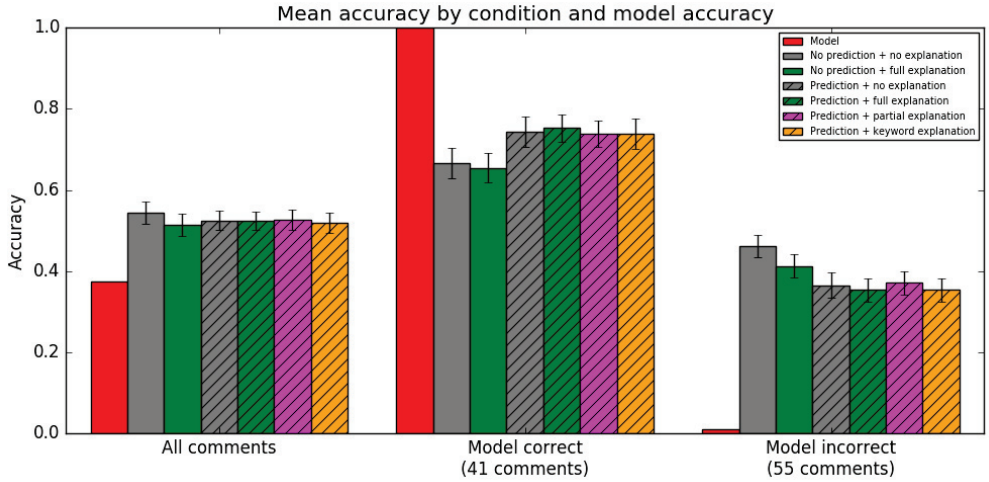

Figure 6: Mean quartile accuracy of model and users across experimental conditions and question subsets with 95% confidence intervals.

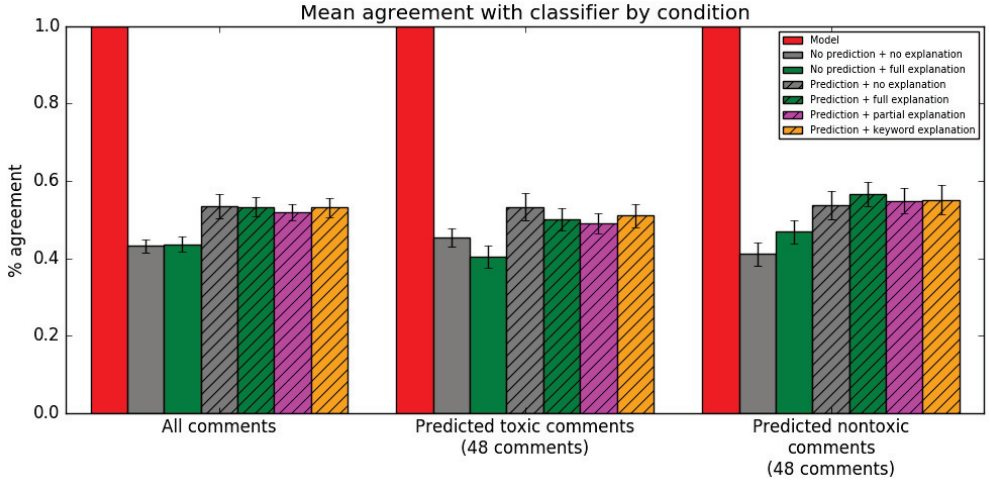

Figure 7: Mean agreement of human subjects with classifier predicted quartile.

This result suggest that in this context, explanations tend to cause subjects to make mistakes of omission rather than interpretation, presumably as they focus on the text that has been highlighted without considering the un-highlighted text (which may sometimes contain evidence of toxicity).

The one unequivocal benefit we do find is that explanations erase the speed penalty of prediction presence, allowing users to more speedily determine whether they believe or disbelieve in the model output.

RQ3: Explanation Type

We do not observe significant differences in outcome among the three explanation variants tested. The "partial" variant performs marginally worse relative the "full" variant on several metrics, but not significantly so. We find no significant difference between the full explanation model and the much simpler "keyword" explanation variant which just highlights potentially problematic words regardless of context.

The comments sampled from the (Wulczyn, Thain, and Dixon 2017) dataset were not optimized for separation between these variants. For example, roughly 50% of comments were predicted-nontoxic, for which the model would have highlighted little or no content across all three variants. Future work investigating the impact of different attribution styles may need to explicitly sample for this type of separation in order observe a difference effect.

Experiment Design for Evaluating Interpretability

The results of this study, in combination with other similar recent studies like (Lai and Tan 2019) and (Nguyen 2018), begin to reveal an expectation gap in the interpretability literature between the excitement that the field has inspired and

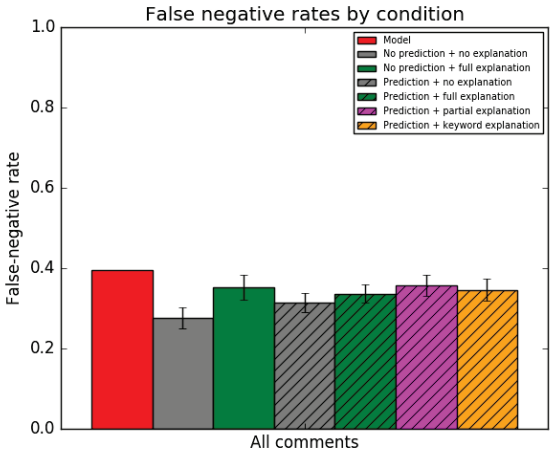

Figure 8: Mean false positive rate of subjects across conditions.

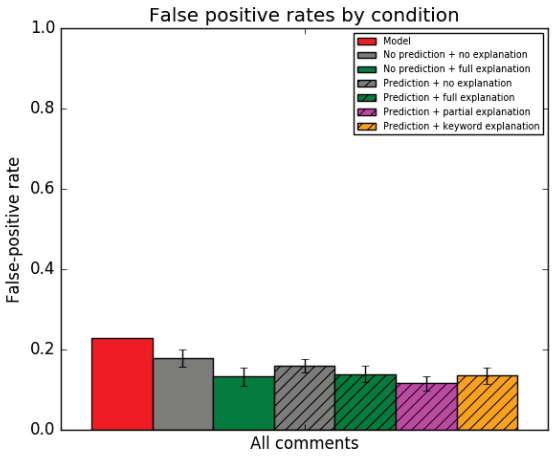

Figure 9: Mean false negative rate of subjects across conditions.

the improvement that the field has been able to demonstrate in human/model performance.

One possible factor in this gap is the relative balance of model and human skillfulness in the design of evaluation studies for explanatory machine learning. In order for explanations to be useful from a decision quality perspective, we argue that they have to allow human operators to make more accurate decisions than either unassisted human baseline accuracy or unsupervised model accuracy. Otherwise, there is no point in combining the two types of agent–one or the other working alone would be a better solution.

For this to be the case, there need to be a substantial proportion of instances for which model performance is good and human performance poor, and vice versa. (Kleinberg et al. 2017) found this to be the case for recidivism prediction, as an example. The more of a performance gulf that exists between baseline human and model performance, the less

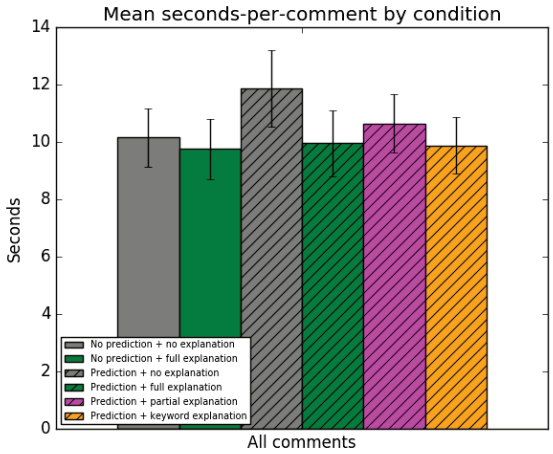

Figure 10: Mean seconds-per-comment of subjects across conditions.

common such instances will be, meaning that the greatest potential improvement exists when human and model baseline performance is roughly equal.

To achieve such a balance in this experiment, we generated a sample of comments from the (Wulczyn, Thain, and Dixon 2017) stratified by model error relative to the existing ground truth, and we were careful to include one experimental condition for assessing baseline human accuracy. While we did not observe an accuracy effect in our study, we believe that failing to account for these issues in future studies may result in spurious effects.

Limitations

The experiment had several limitations that would have to be addressed in further work. First, the three explanation variants we test are not fully representative of the current interpretability literature. Rather, they represent three extremes: capturing *all* locally pertinent information (full variant), capturing *minimal* locally pertinent information (partial variant), and capturing independent globally pertinent information (keyword variant). It is possible that there exists some feature highlighting technique (e.g. (Arras et al. 2017)) that would produce better outcomes, though it seems unlikely given the relatively consistent negative result across the three explanation variants.

We also limited ourselves to discrete binary highlighting–a token is either in or out of an explanation, without further embellishment. We did not include words and phrases of nontoxic valence, nor did we allow for grades of relevance, as in (Arras et al. 2017). It is possible that a more informative style of feature-highlighting would produce the accuracy benefits that we failed to observe in this work.

We displayed all information at once–that is, text, prediction and highlighting were all presented together to each user. A multi-phase presentation, where users are prompted for an initial decision before being exposed to any algorithmic assistance, might result in less bias toward the model prediction. However, it would also reduce the potential for

time savings, as users would have to go to all the trouble of making a careful decision before getting a chance to process the output of the algorithm.

Finally, the stratified question sets we employed in this experiment are significantly more toxic than a random sample of social media comments would be, while the model was significantly less accurate than a model would be on randomly distributed data (38% versus the 86%). While we were able break down task performance by individual-comment accuracy, the particular distribution of toxicity and classifier error probably prompted subjects to be more skeptical of the model and differently sensitive to toxicity than if the comments had been sampled in a more representative manner.

Toxicity Detection Versus Moderation

Our experiment involves untrained Mechanical Turk workers making predictions about the consensus toxicity of Wikipedia revision comments. The comments they view represent a variety of different true levels of toxicity and are removed from conversational context. This is quite abstracted from a true moderation setting, where trained moderators apply a specific set of community standards to comments, typically in response to some kind of reporting mechanism.

However, the purpose of this study is less to prototype a model-assisted moderation system than to test the impact of interpretability on model-assisted human performance on a decision task that involves a tension between existing intuitions and an external standard for correctness. In a true moderation task, the external standard would consist of a set of community guidelines; in our experiment it is the consensus label established by the Phase 1 labeling task.

The question of the difference between "toxicity detection" and moderation is an important one, but it is also one that belongs to the larger literature on machine approaches to online abuse. The Perspective API and the response it has generated from the research community represent some of the dialogue surrounding this question, though we are not aware of an existing theoretically-motivated attempt specifically to reconcile the toxicity detection task with the task of moderation as experienced by real-world moderators.

Design Implications

This study has several implications for systems which seek to provide advice from a text classifier to a human worker, particularly on a subjective task such as toxicity detection.

First, our results suggest that feature highlighting is not sufficient by itself to improve human accuracy on ambiguous items, possibly because it doesn't do enough to clarify the relationship between the highlighted features and the suggested target value. Other types of explanations may be needed in order to achieve any improvement on this front.

Furthermore, the way in which explanations change the distribution of human error suggests that a system builder needs to be very careful in their choice of explanation mechanism, because explanations could exacerbate a pre-existing tendency toward false negatives. If time is not a factor, it may actually be better in some cases to have no explanatory mechanism, as this forces users to be thorough in resolving any disagreement between themselves and the advisory classifier. The good news is that the lack of difference between the full and keyword-based explanation types suggest that simple highlighting methods, even dictionary methods, can be just as effective as more sophisticated ones, as long as they are tuned to catch as much relevant content as possible.

Finally, the high-level implication of this study and others like it is that interpretable machine learning is not necessarily the panacea to unreliable models that the interpretability literature tends to assume it is–the most popular type of explanation fails to improve how well humans use such a model, and poor explanations can actually reduce human performance by discouraging critical thinking about the model's predictions.

Conclusion

In this study we test how the presence of feature attribution-style explanations for a text toxicity classifier impact the ability of humans to make effective use of that classifier. We find that such explanations do not improve accuracy or trust in the classifier, though they do remove the speed penalty associated with the presence of model advice. Explanations also have a tendency to increase false negatives while decreasing false positives. We find no significant difference in outcomes between three variant attribution methods. Our study has implications for the design of systems which seek to combine classifier and human effort–it suggests that more sophisticated and informative types of explanations may be needed to improve human accuracy on these types of joint classification tasks.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant numbers 1717688, 1633370 and 1620319. We thank Eric Gilbert, Nicole Ellison and Daphne Chang for their contributions to the development of the paper.

References

Abdul, A.; Vermeulen, J.; Wang, D.; Lim, B. Y.; and Kankanhalli, M. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18. Montreal QC, Canada: ACM Press.

Anderson, A. A.; Yeo, S. K.; Brossard, D.; Scheufele, D. A.; and Xenos, M. A. 2016. Toxic Talk: How Online Incivility Can Undermine Perceptions of Media. *International Journal of Public Opinion Research*.

Arras, L.; Horn, F.; Montavon, G.; Mller, K.-R.; and Samek, W. 2017. "What is relevant in a text document?": An interpretable machine learning approach. *PLOS ONE* 12(8):e0181142.

Bansal, G.; Nushi, B.; Kamar, E.; Weld, D. S.; Lasecki, W. S.; and Horvitz, E. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33:2429–2437.

Binns, R.; Veale, M.; Van Kleek, M.; Shadbolt, N.; Veale, M.; Van Kleek, M.; and Shadbolt, N. 2017. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In *Social Informatics*, volume 10540, 405–415. Cham: Springer.

Blackwell, L.; Chen, T.; Schoenebeck, S.; and Lampe, C. 2018a. When Online Harassment is Perceived as Justified. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*.

Blackwell, L.; Dimond, J.; Schoenebeck, S.; and Lampe, C. 2018b. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2018)*, 1–19.

Carton, S.; Mei, Q.; and Resnick, P. 2018. Extractive Adversarial Networks: High-Recall Explanations for Identifying Personal Attacks in Social Media Posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3497–3507. Brussels, Belgium: Association for Computational Linguistics.

Chancellor, S.; Kalantidis, Y.; Pater, J. A.; De Choudhury, M.; and Shamma, D. A. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 3213–3226. New York, NY, USA: ACM.

Chandrasekharan, E.; Samory, M.; Jhaver, S.; Charvat, H.; Bruckman, A.; Lampe, C.; Eisenstein, J.; and Gilbert, E. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction - CSCW* 2(CSCW):1–25.

Chandrasekharan, E.; Gandhi, C.; Mustelier, M. W.; and Gilbert, E. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–30.

de Gibert, O.; Perez, N.; Garca-Pablos, A.; and Cuadros, M. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online*, 10.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2019. ERASER: A Benchmark to Evaluate Rationalized NLP Models. *arXiv preprint*. arXiv: 1911.03429.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, 67–73. New York, NY, USA: ACM. event-place: New Orleans, LA, USA.

Doshi-Velez, F., and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint*. arXiv: 1702.08608.

Dove, G.; Halskov, K.; Forlizzi, J.; and Zimmerman, J. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 278–288. Denver, Colorado, USA: ACM Press.

Eslami, M.; Krishna Kumaran, S. R.; Sandvig, C.; and Karahalios, K. 2018. Communicating Algorithmic Process in Online Behavioral Advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 432:1–432:13. New York, NY, USA: ACM. event-place: Montreal QC, Canada.

Feng, S.; Wallace, E.; Grissom II, A.; Iyyer, M.; Rodriguez, P.; and Boyd-Graber, J. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Fiesler, C.; Jiang, J. A.; McCann, J.; Frye, K.; and Brubaker, J. R. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Twelfth International AAAI Conference on Web and Social Media*.

Fortuna, P., and Nunes, S. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys* 51(4):1–30.

Friedler, S. A.; Roy, C. D.; Scheidegger, C.; and Slack, D. 2019. Assessing the Local Interpretability of Machine Learning Models. *arXiv preprint*. arXiv: 1902.03501.

Gillies, M.; Lee, B.; d'Alessandro, N.; Tilmanne, J.; Kulesza, T.; Caramiaux, B.; Fiebrink, R.; Tanaka, A.; Garcia, J.; Bevilacqua, F.; Heloir, A.; Nunnari, F.; Mackay, W.; and Amershi, S. 2016. Human-Centred Machine Learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*, 3558–3565. Santa Clara, California, USA: ACM Press.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *arXiv preprint*. arXiv: 1806.00069.

Golbeck, J.; Gnanasekaran, R. K.; Gunasekaran, R. R.; Hoffman, K. M.; Hottle, J.; Jienjitlert, V.; Khare, S.; Lau, R.; Martindale, M. J.; Naik, S.; Nixon, H. L.; Ashktorab, Z.; Ramachandran, P.; Rogers, K. M.; Rogers, L.; Sarin, M. S.; Shahane, G.; Thanki, J.; Vengataraman, P.; Wan, Z.; Wu, D. M.; Banjo, R. O.; Berlinger, A.; Bhagwan, S.; Buntain, C.; Cheakalos, P.; Geller, A. A.; and Gergory, Q. 2017. A Large Labeled Corpus for Online Harassment Research. In *Proceedings of the 2017 ACM on Web Science Conference*, 229–233. ACM Press.

Hosseini, H.; Kannan, S.; Zhang, B.; and Poovendran, R. 2017. Deceiving Googles Perspective API Built for Detecting Toxic Comments. *arXiv preprint*. arXiv:1702.08138.

Hosseinmardi, H.; Mattson, S. A.; Ibn Rafiq, R.; Han, R.; Lv, Q.; and Mishra, S. 2015. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. In Liu, T.-Y.; Scollon, C. N.; and Zhu, W., eds., *Social Informatics*, Lecture Notes in Computer Science, 49–66. Springer International Publishing.

Jain, S., and Wallace, B. C. 2019. Attention is not Explanation. *arXiv preprint*. arXiv: 1902.10186.

Jhaver, S.; Birman, I.; Gilbert, E.; and Bruckman, A. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26(5):31:1–31:35.

Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2017. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*.

Kumar, R.; Ojha, A. K.; Zampieri, M.; and Malmasi, S. 2018. Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S.; and Doshi-Velez, F. 2019. An Evaluation of the Human-Interpretability of Explanation. *arXiv preprint*. arXiv: 1902.00006.

Lai, V., and Tan, C. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 17.

Lakkaraju, H.; Kamar, E.; Caruana, R.; and Leskovec, J. 2019. Faithful and Customizable Explanations of Black Box Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 131–138. Honolulu HI USA: ACM.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 107–117.

Lundberg, S. M., and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* 10.

Malmasi, S., and Zampieri, M. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence* 30(2):187–202.

Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; and Yu, B. 2019. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint*. arXiv:1901.04592.

Napoles, C.; Tetreault, J.; Pappu, A.; Rosato, E.; and Provenzale, B. 2017. Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*.

Nguyen, D. 2018. Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1069–1078.

Olteanu, A.; Talamadupula, K.; and Varshney, K. R. 2017. The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection. In *Proceedings of the 2017 ACM on Web Science Conference - WebSci '17*, 405–406. Troy, New York, USA: ACM Press.

Pavlopoulos, J.; Malakasiotis, P.; and Androutsopoulos, I. 2017. Deep Learning for User Comment Moderation. In *Proceedings of the First Workshop on Abusive Language Online*, 25–35.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Vaughan, J. W.; and Wallach, H. 2018. Manipulating and Measuring Model Interpretability. *arXiv preprint*. arXiv: 1802.07810.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 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. ACM.

Salminen, J.; Hopf, M.; Chowdhury, S. A.; Jung, S.-g.; Almerekhi, H.; and Jansen, B. J. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences* 10(1):1.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. Florence, Italy: Association for Computational Linguistics.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features Through Propagating Activation Differences. *arXiv preprint*. arXiv: 1704.02685.

Springer, A.; Hollis, V.; and Whittaker, S. 2017. Dice in the Black Box: User Experiences with an Inscrutable Algorithm. In *2017 AAAI Spring Symposium Series*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. *arXiv preprint*. arXiv: 1703.01365.

Svec, A.; Pikuliak, M.; Simko, M.; and Bielikova, M. 2018. Improving Moderation of Online Discussions via Interpretable Neural Models. In *Proceedings of the 2nd Workshop on Abusive Language Online*.

Wang, C. 2018. Interpreting Neural Network Hate Speech Classiers. In *Proceedings of the 2nd Workshop on Abusive Language Online*, 7.

Waseem, Z.; Chung, W. H. K.; Hovy, D.; and Tetreault, J., eds. 2017. *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC, Canada: Association for Computational Linguistics.

Weerts, H. J. P.; van Ipenburg, W.; and Pechenizkiy, M. 2019. A Human-Grounded Evaluation of SHAP for Alert Processing. *arXiv preprint*. arXiv: 1907.03324.

Wiegreffe, S., and Pinter, Y. 2019. Attention is not not Explanation. *arXiv preprint*. arXiv: 1908.04626.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399.

Yu, M.; Chang, S.; Zhang, Y.; and Jaakkola, T. S. 2019. Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control. *arXiv preprint*. arXiv: 1910.13294.