Using Off-the-Shelf Harmful Content Detection Models: Best Practices for Model Reuse

ANGELA SCHÖPKE-GONZALEZ, University of Michigan School of Information, USA
SIQI WU, Indiana University Bloomington, USA
SAGAR KUMAR, Northeastern University, USA
LIBBY HEMPHILL, University of Michigan School of Information, USA

Supervised machine learning is a common approach for automated harmful content detection to support content moderation. This approach relies on data annotated by humans to train models to recognize classes of harmful content. For detection tasks, researchers or content moderation communities typically either design their own annotation tasks to generate training data for new harmful content detection models, or use off-the-shelf (OTS) pre-trained harmful content detection models. OTS model reuse can enable detection tasks in resource-constrained contexts and can help to reduce the environmental impact of training new models – an energy-intensive process. However, given the plethora of OTS models now available for reuse, determining which OTS model to reuse for a particular task and how to use it can be challenging, especially given that many of these models have been developed for specific contexts that are not always easily transferred onto others. This work aims to provide best practices for reusing OTS models for harmful content detection tasks. By using content analysis and statistical methods to evaluate assumptions about OTS model utility and reusability, we show that model reusers cannot assume that a model claimed to detect a particular concept, will actually detect that concept. Instead, based on our findings, we offer a decision tree for how to assess whether an OTS model would be appropriate for reuse for a new harmful content detection task. This decision tree directs model reusers to critically assess concept definitions, annotation task design, and additional features specified in our content analysis codebook to identify expected model output, and consequently evaluate whether that OTS model is appropriate for reuse for a new detection task.

CCS Concepts: • **Applied computing** → *E-government*; • **General and reference** → **Computing standards, RFCs and guidelines**; *Evaluation.*

Additional Key Words and Phrases: Model Reuse, Content Moderation, Machine Learning, Content Analysis

ACM Reference Format:
Angela Schöpke-Gonzalez, Siqi Wu, Sagar Kumar, and Libby Hemphill. 2025. Using Off-the-Shelf Harmful Content Detection Models: Best Practices for Model Reuse. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW201 (April 2025), 27 pages. https://doi.org/10.1145/3711099

1 INTRODUCTION

The volume and velocity of content generated by social media can be challenging for moderation workers to keep up with. Moderation work can also contribute to severe psychological distress for workers [44, 50]. Some extent of automated or computational assistance can support content moderation by reducing workload or filtering some psychologically distressing content. Supervised machine learning (ML) is a common tool used to automatically detect harmful content. Supervised

Authors' addresses: Angela Schöpke-Gonzalez, aschopke@umich.edu, University of Michigan School of Information, 105 South State Street, Ann Arbor, Michigan, USA, 48109-1285; Siqi Wu, Indiana University Bloomington, Bloomington, Indiana, USA; Sagar Kumar, Northeastern University, Boston, Massachusetts, USA; Libby Hemphill, University of Michigan School of Information, 105 South State Street, Ann Arbor, Michigan, USA, 48109-1285.

ACM 2573-0142/2025/4-ARTCSCW201
https://doi.org/10.1145/3711099

ML relies on data annotated by humans to train models to recognize classes of harmful content. For detection tasks, researchers or content moderation communities typically either design their own annotation tasks to generate training data for new harmful content detection models, or use off-the-shelf (OTS) pre-trained[1] harmful content detection models. Some examples of OTS models reused for harmful detection tasks include Jigsaw's Perspective API [28], Amazon Rekognition [4], Azure's Content Moderation API [30], WebPurify API [36], and DeepAI's content moderation API [18]. As a practice, OTS model reuse can enable detection tasks in resource-constrained contexts, for example when researchers or content moderation communities do not have access to the computing power, people, time, and funding necessary to train new models. Furthermore, reusing OTS models can help to reduce the environmental impact of training new models, an energy-intensive process [48]. However, given the plethora of OTS models now available for reuse, determining which OTS model to reuse for a particular task and how to use it can be challenging, especially since many of these models have been developed for specific uses that are not easily mapped to other contexts [52].

Building on the lineage of Vidgen and Derczynski [56]'s best practices for designing training datasets and Pustejovsky and Stubbs [42]'s best practices for developing annotation guidelines for training datasets, this paper aims to provide best practices for reusing pre-trained OTS models for harmful content detection tasks. We achieve this goal by first identifying how model reusers can make sense of OTS harmful content detection models' outputs, and second proposing a decision tree model for steps that model reusers can take to determine whether and how to reuse these OTS models.

We might assume that OTS harmful content detection models are useful and can be reused to identify those harm concepts that models are designed to detect. For example, we might assume that a model designed to identify 'hateful' content is useful and can be reused for identifying 'hateful' content. To test this assumption, we evaluate a dataset that simulates the re-creation of training datasets for three common OTS harmful content detection models [60]. By using content analysis and statistical methods to evaluate assumptions about OTS model utility and reusability, we show that model reusers cannot assume that a model is internally valid and that it will actually detect the concept that it is designed to identify. Instead, we find that features including harm concept definitions, tone, sarcasm, who comments are directed at, insults, name-calling, and annotators' subjective interpretive lenses all affect labeling outcomes. We provide a decision tree of best practices for OTS model reuse directing model reusers to assess concept definitions, annotation task design, and additional features specified in our content analysis codebook to identify expected model output, and determine whether that OTS model is appropriate for reuse for their new detection task.

2 BACKGROUND & LITERATURE REVIEW

As researchers have grappled with what constitutes harm and how to automatically detect it in text corpora, they have generated models intended to be reusable OTS. Among these OTS models popular for reuse are Jigsaw's Perspective API used or discussed in over 750 research publications according to a Google Scholar search in July 2024 [28], Davidson et al.'s [10] model to detect offensiveness cited over 2900 times according to the same Google Scholar search, and Aluru et al.'s [2] model for detecting hate speech cited 36 times. Examples of model reuse in research include model replication and testing [21], using OTS models trained on English-language corpora as frameworks or baselines for non-English language models [51], and studying social phenomena like

[1]We use the expressions "OTS" and "pre-trained" interchangeably to address and engage with different naming conventions across researchers and content moderation professionals who may be interested in model reuse.

the effects of shadowbans Twitter users' behaviors [26] or what kinds of news articles generate the most hateful comments [62]. Outside of research contexts, Perspective API has been used to support harmful content detection in the comments sections of online news publications like The New York Times, El País, Le Monde, The Financial Times, Southeast Missourian, and several others [5]. Also outside of research contexts, Aluru et al.'s model has been used to develop a web application [16] that identifies hateful content in Portuguese electoral manifestos [15]. While other OTS models are available, we orient our literature review around these three models as case studies given their popularity and their relevance to the dataset that we analyze [60]. The following subsections describe how these models came to be, and how we can understand these OTS models' utility for reuse.

2.1 Emergence of Reusable OTS Harmful Content Detection Models

In the late 2010s, Hatebase.org – a collaboration between Dark Data Project [40] and The Sentinel Project [41] – curated a then popularly used lexicon of 'offensive' language that social media platform users often perceived as 'hate' speech [24]. Hatebase.org characterized offensiveness as a dimension of hate rather than as a distinct, standalone concept. While Hatebase.org did not explicitly define offensiveness, Wiegand et al.'s [59] later definition captures a similar spirit: offensiveness can in part be identified according to the content of phrases that the comment uses (i.e., "obscene"), similar to Hatebase.org's lexicon. While Hatebase.org characterized offensiveness as a subset of hate, when Davidson et al. [10] investigated differences between hate speech and offensive language, they found that only about 5% of content containing offensive language was interpreted as hate speech by human annotators. They find that offensiveness is not a good proxy for hate, nor is it always a dimension or subset of hate. Based on these findings, Davidson et al. [10] trained a model for detecting offensiveness as distinct from hate. This model is now an available for others to reuse [11].

Much like Hatebase.org did not define offensiveness explicitly, when labeling a training dataset for detecting offensiveness with their model, Davidson et al. [10] did not define offensiveness. Rather, they provided annotators with a definition for hate speech and subsequently asked crowdworkers to annotate data as either hate speech, offensive but not hate speech, or neither offensive nor hate speech. Davidson et al.'s hate speech distinguished itself from Hatebase.org's and Wiegand et al.'s [59] offensiveness – characterized by comment content – by explicitly focusing on expressions *directed at* groups or members of a group, and content that the comment author *intended* to be "derogatory, to humiliate, or to insult" [10]. However, Davidson et al.'s hate speech definition was only one of many at the time. Additional definitions like Nockleby's by way of Schmidt and Wiegand's [49] suggest that hate speech can also encompass expressions *directed at* individual persons or groups, and comments with particular *effects* on their recipients (i.e., disparaging). ElSherief, Kulkarni, et al. [13] and ElSherief, Nilizadeh, et al. [14] further specify that hate speech encompasses only expressions *directed at* an individual person that have particular *effects* on that person (i.e., denigrates). Even before Davidson et al. had appeared on the scene, Waseem and Hovy's [58] definition clearly defined specific *behaviors* – use of a slur, attacking, seeking to silence, criticizing, misrepresenting, etc. – that when used in combination with *directedness* toward an individual or group and, in some cases, with particular author *intents*, manifest hate speech. Consolidating rapidly proliferating definitions of 'hate', Aluru et al. [2] trained a harmful content detection model – now an OTS model available for others to reuse [3] – on a combination of training datasets annotated with different 'hate' definitions and across multiple languages.

Around the same time that Davidson et al.'s released their model, Jigsaw – a unit within technology company Google – released its now popular Perspective API, a harmful content detection model trained to detect 'toxic' content [28]. Jigsaw's definition of toxicity differs from hate speech

and offensive language in that it does not refer to qualities of a comment like *directedness*, *behaviors*, or *specific language*, but rather only describes a comment's *effect*. A later effort to define toxicity in the context of gaming communities by Beres and colleagues [7] similarly defines toxicity according to its *effects* on gameplay. Although at first glance perhaps similar to Davidson et al.'s offensiveness detection model or Aluru et al.'s hate detection model, Perspective API differed considerably in its approach to identifying harm.

2.2 Relationships between Concept Definitions and Pre-trained Model Utility

When designing annotation tasks for training datasets, guidelines recommend defining concepts as specifically as possible to increase the consistency of interpretation and thus the consistency of the labels between annotators [42]. None of those definitions used to train Davidson et al.'s, Aluru et al.'s, and Jigsaw's original OTS models follows this guidance, making their definitions ambiguous and their utility for identifying specific harm concepts questionable. In effect, a model *designed* to detect 'hateful' content may not actually be that useful for detecting 'hateful' content. Instead, these models may be useful for identifying generalized harm concepts predicted by an annotator's subjective idea about anticipated *effect* of a comment on an audience (e.g., Jigsaw's toxicity), or that are *directed at* some entity (e.g., many of the hate definitions used to develop the combined dataset Aluru et al.'s trained their model on).

To address these concerns about harm detection models' internal validity and thus generalizability, prior work has explored various steps to enhance these models' generalizability. For instance, Jin and colleagues [29] experimented with procedures for improving model transfer by using bias mitigation algorithms to fine-tune models' "upstream" encoding tasks – tasks that affect a model's ability to interpret and classify text – and then using that fine-tuned encoder for "downstream" classification tasks. Their "upstream bias mitigation" approach effectively improved cross-task model performance and may be a useful approach for adapting models trained to detect similar, but not identical concepts. However, many OTS models do not provide direct access to their encoders or classifiers, making using such approaches to improve their performance impossible. Additional approaches to fine-tuning focus on debiasing datasets themselves through techniques like data augmentation [12] or adversarial learning [35]. However, like for Jin et al.'s upstream bias mitigation, while some OTS models do provide direct access to their original training datasets to apply these techniques, others do not, making the application of these techniques impossible. Like many natural language processing tasks, harmful content detection has turned to large language models (LLMs) for improvement. However, LLMs are usually trained on large, general datasets, and their performance degrades as tasks get more domain specific. Researchers have pointed out that fine-tuning these models does not effectively improve their performance [23]. Instead, multiple, domain-specific pre-training steps are required [22]. To achieve these improvements, model reusers must have the resources to run existing models, collect appropriate training data for fine-tuning, and conduct fine-tuning. Many social scientists, content moderators, or other OTS model reusers lack the necessary computational and financial resources for these steps.

In parallel, several scholars have developed methods for addressing internal validity issues that do not require access to encoders, classifiers, and training datasets. These efforts target a particular threat to internal validity: misclassification errors. Some examples of methods that aim to improve internal validity through addressing misclassification errors include Fong et al.'s general method of moments estimator [17], Blackwell et al.'s multiple imputation framework [8], and TeBlunthuis et al.'s maximum likelihood adjustment method [54]. However, as TeBlunthuis et al. point out, these methods require making an assumption that the human annotations used to train models are error-free, which may not always be the case, especially in the event of ambiguous harmful concept definitions.

Given limitations and constraints of these efforts to address OTS model generalizability for harmful content detection and OTS models' internal validity challenge, we investigate which features *do* predict model labels when harmful concept definitions are ambiguous. In the following section, we identify features that we anticipate may predict harmful labels, our hypotheses, and methods that we use to test our hypotheses.

3 METHODS

In order to assess the utility of three OTS harmful content detection models, we use a dataset [60] that simulates the re-creation of training datasets for Aluru et al.'s hate (H) detection model [2], Davidson et al.'s offensiveness (O) detection model [10], and Jigsaw's toxicity (T) detection model [28]. We identify potentially predictive features of whether a comment will be labeled as any of H, O, or T by reviewing harm concept definitions and using content analysis. We then establish hypotheses for how we expect these features to predict labels, and perform analyses to assess our hypotheses.

3.1 Dataset

Wu et al.'s [60] dataset includes annotations by MTurk workers identifying which, if any, of concepts H, O, and T applies to 3481 social media comments posted in response to political news posts and videos on Reddit, Twitter, and YouTube in August 2021. While the dataset's annotation task does use Jigsaw's T definition in its re-creation of labeled datasets, since Davidson et al. did not define O in their annotation task, Wu et al. used Wiegand et al.'s related definition for O [59]. Similarly, since Aluru et al.'s H detection model is trained on an amalgamation of datasets that define H differently, Wu et al. used Davidson et al.'s H definition [10]. See Table 1 for an overview of these definitions.

Wu et al.'s sample of 3481 social media comments was purposefully sampled from a larger dataset of comments to produce a balance of approximately half of all comments anticipated to be H, O, or T (assessed using Jigsaw, Davidson et al.'s and Aluru et al.'s models' predicted labels for the dataset), and the other half anticipated *not* to be H, O, or T [60]. In this dataset, each comment received annotations from five MTurk workers.[2] Each worker assessed whether each of the three concepts applied to the comment. In most of our analyses, we aggregate the annotations from the five MTurk workers, making a majority vote label for each of the concepts. If a comment received three or more annotations indicating the presence of a concept, the comment received a *true* value for that concept. If it received two or fewer annotations indicating the presence of a concept, the comment received a *false* value for that concept. Additional details about the dataset and data collection process are available from the archived dataset record,[3] and we provide a sample of majority-labeled comments in Table 9 in the Appendix.

3.2 Identifying Comment Features

We identified comment features that may predict H, O, or T labels in the Wu et al. dataset by reviewing harm concept definitions and using content analysis. In our review of harm concept definitions (see Appendix Table 5 for full table of harm concept definitions evaluated), harm definitions tend to vary across three key dimensions: comment-based features (*directedness*, *behaviors*, *tone*, or *specific language*), author-based features (*intention*), and audience-based features (*effect*). Figure 1 visualizes these features, and Table 1 dissects Davidson et al.'s hate [10], Wiegand et al.'s offensiveness [59],

[2]Prior to annotating this dataset, each MTurk worker was asked to annotate three qualification comments. If their labels were the same as Jigsaw, Davidson et al.'s and Aluru et al.'s models' predicted labels (validated by Wu et al.), the MTurk worker passed the qualification task and was invited to participate in Wu et al.'s dataset annotation [60].
[3]https://doi.org/10.3886/45fc-9c8f

and Jigsaw's toxicity [28] definitions – definitions used by Wu et al. in annotation instructions – according to this anatomy.

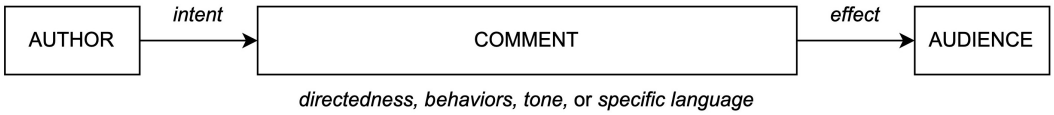

Fig. 1. Anatomy of a comment according to the authors' review of various harm concept definitions (refer to Table 5 in Appendix for full collection of harm concepts reviewed).

Table 1. According to anatomy of a comment in Figure 1, dissected definitions of "hateful", "offensive", and "toxic" that were provided to MTurk workers in Wu et al.'s [60] instructions about how to annotate their collection of Reddit, Twitter, and YouTube comments.

Concept	Definition
Hateful	"expresses hatred [author-based *intent*] towards a **targeted group** [comment-based *directedness*] or is intended to be derogatory, to humiliate, or to insult [author-based *intent*] the **members of the group** [comment-based *directedness*]." [10]
Offensive	"contains *hurtful, derogatory,* [audience-based *effect*] or **obscene comments** [comment-based *specific language*]." [59]
Toxic	"a *rude, disrespectful,* [audience-based *effect*] or unreasonable [author-based *intent*] that *is likely to make readers want to leave a discussion* [audience-based *effect*]."[28]

In addition to using definitions of harm to identify potential predictive features, we used content analysis [34] to determine whether comment features *not* explicitly identified by harm concept definitions may predict whether a comment was majority labeled as H, O, or T. We developed our own codebook using a directed coding approach [25] on a sample of data (220 comments), and applied this codebook to a larger sample (909 comments including the original 220). While our final codebook does overlap with concept definition features, Goyal et al.'s [20] specialized rater pools codebook, and Waseem et al.'s [57] proposed typology of harmful content assessment, for example, developing our own codebook allowed us to make space for additional features not captured by concept definitions or existing codebooks.

Before beginning codebook development, three authors met to discuss how to navigate reviewing content that may cause emotional distress. Strategies included taking breaks and ensuring that we had timespace after reviewing content to externalize anger or hopelessness. We noted differences in our life experiences that could shape how we interpret content. For example, one researcher's relationship with immigration, Afghanistan, and gender-based violence meant that she was particularly attuned to nuances in texts concerning these topics. The second researcher's experiences studying white supremacy and homophobia made her aware of references to religious, racial, and sexual minorities. The third researcher's previous personal and professional work unpacking ideas of "conspiracy" or "misinformation" attuned him to nuances in texts related to these topics.

To develop our codebook, we randomly selected 120 comments that any annotator had indicated were H, O, and/or T. Two of the researchers independently used directed coding [25] to code the first 60 of the 120 comments, and then met to discuss emergent patterns they noticed (see Table 8 in the Appendix for a sample coded comment). The two researchers codified emergent patterns under three feature types that may predict H, O, and T labels – *directedness*, *pronoun use*, and

behavior. The two researchers then independently reviewed the remaining 60 comments of the random sample. Once the two authors had coded this 120-comment sample, one of the researchers, together with the third researcher, coded an additional 100 randomly selected comments to assess whether the codes developed by the first two researchers were transferable to a third researcher. All three researchers met to discuss patterns they noticed. While the third researcher's resulting features overlapped with the first two researchers' three identified features, discussion led to their disentanglement from two additional feature types – *communication style* and *sarcasm*. In instances of disagreement in how we applied features, we reached resolution by reviewing feature definitions, refining definitions if necessary, and reassessing how our applications compared to the refined definition.

In total, we identified five comment features (see Table 6 for feature descriptions and examples) that may predict H, O, and T labels: *directedness, pronoun use, behavior, communication style,* and *sarcasm*. Some of these features overlap with definitional features reviewed in section 2, specifically: *directedness*, who or what a comment is directed at, and type of *behavior* exhibited by the comment author [7, 58]. *Communication style* and *sarcasm* approximate tone as a feature [9], and *pronoun use* and *multiple behaviors* are codes that newly emerged from the data. In addition, some of these features overlap with existing codebooks like Goyal et al.'s [20] specialized rater pools codebook, which includes *insult* and *threat* as parent categories describing behaviors. In research concerning ethnic bias of language models, Jeong et al. [27] also identified targetedness (what we call *directedness*) of a comment at a group or individual as a feature of what characterizes offensiveness in Korean language posts. *Directedness* (as opposed to "generalizedness") of a comment is also a feature in Waseem et al.'s [57] proposed typology of harmful content assessment. In addition to these features, our codebook encompasses a wider range of features that may predict H, O, and T labels.

3.3 Hypotheses

Based on H, O, and T definitions that Wu et al. provided to annotators and those potentially predictive comment features we identified, we developed hypotheses for which features we expect to be predictive of whether annotators labeled a comment as any of H, O, or T. Of Davidson et al.'s, Wiegand et al.'s, and Jigsaw's harm concept definitions, the first two of these definitions appear to attend best to Pustejovsky and Stubbs' [42] guideline that annotation tasks clearly define concepts. Wiegand et al. articulate a specific feature of a comment ('obscenity') that an annotator can identify as offensive, and Davidson et al. articulate *directedness* at a group of people as a comment feature identifiable as hateful. Accordingly, we hypothesize that the more specifically a harm concept definition directs annotators to identify specific comment features as a marker of harm, the more consistent annotations will be across annotators. In particular:

> **H1:** H and O concepts will experience more consistent annotations across annotators than T.

We also anticipate that features of comments labeled with a particular harm concept will reflect features indicated in the concept definition. Specifically:

> **H2:** *Directedness* at groups will be a predictive feature of comments labeled as H.

Since O's specified comment feature is specific language – 'obscene comments' – but obscenity is not defined and thus more subject to individual raters' biases,[4] we do not expect that this comment feature will be predictive of O labels. We call 'obscene comments' a nested term, or a term that is embedded in another term's concept definition. In the absence of specifically articulated comment

[4]Though the term 'bias' is often used to characterize error, we instead understand bias as the complex and entangled factors that shape how individuals uniquely perceive and interpret the world.

features associated with harm concepts, we expect that comments' harm concept labels will be predicted by comment features not made explicit in concept definitions. Specifically:

> **H3:** All comments will be predicted by comment features outside of those made explicit in concept definitions.

Finally, additional components of Wiegand et al.'s and Davidson et al.'s harm definitions, and Jigsaw's definition in its entirety, attend only to author-based and audience-based features that annotators do not have information about when annotating training data. In the absence of this information, we expect that individual raters' biases will be more important for a rater in determining whether a comment is harmful or not. Accordingly, we hypothesize that the less specific a harm concept definition is in directing annotators to identify specific comment features as a marker of harm, the more likely individual raters' biases will affect annotation outcomes. In particular, since Jigsaw's T definition does not direct annotators to identify specific comment features as markers of harm, we expect that

> **H4:** Individual raters' biases will be more predictive of T label outcomes than either of H and O label outcomes.

3.4 Krippendorff's Alpha & Statistical Analyses

In order to evaluate **H1**, we used Krippendorff's alpha [32] – a measure for interannotator reliability. Since in Wu et al.'s data each comment only received ratings from five independent randomly selected crowdworkers and the worker pool consisted of 608 crowdworkers, each comment experiences 603 missing annotations. Krippendorff's alpha was thus the most appropriate interannotator reliability measure for our analysis because of its stability if data are missing [63].

In order to evaluate hypotheses **H2**, **H3**, and **H4**, we used generalized linear regression models – an appropriate method for data where outcome variables are binary – through R's stats package [43] using comment features as predictors, and H, O, and T labels as outcome variables. These models can be described by the general equation:

$$Pr(y = 1) = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \alpha_i)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \alpha_i)} \tag{1}$$

where y denotes the binary outcome variable for each concept, x denotes each independent variable, and α_i denotes a fixed effect for an annotator group. Since we randomly assigned a group of five annotators from our annotator pool to each social media comment, we needed to account for the effect on labeling of both the individual annotator and the group of five annotators that labeled each comment (i.e., annotator group). To account for the effect of each individual annotator on a group's likelihood to label a comment with each concept, we calculated a fixed effect α_i for each comment i by concept where j denotes an annotator; x denotes a label (0 or 1); k denotes a single comment that an annotator has labeled; and n denotes the total number of comments an annotator has labeled:

$$\alpha_i = \frac{\sum_{j=1}^{5} (\frac{\sum_{k=1}^{n} x_{k,j}^{I})}{n}}{5} \tag{2}$$

Since the features *directed at* and *pronoun use* apply only to directed comments, we separated the data into subsamples of directed comments (823 comments total) and undirected comments (87 comments total). Given undirected comments' small sample size, we statistically analyzed only directed comments. Models with the directed comment subset included predictors *directed at, pronoun use, behavior, communication style, sarcasm,* and *rater group effect.*

Table 2. Variable definitions.

Outcome Variable	Variable Definition	Count TRUE	% TRUE
H (all)	All comments labeled hateful.	148	17.98%
O (all)	All comments labeled offensive.	281	34.14%
T (all)	All comments labeled toxic.	288	35.00%
H (only)	Comments labeled only hateful, but not offensive or toxic.	21	2.55%
O (only)	Comments labeled only offensive, but not hateful or toxic.	59	7.17%
T (only)	Comments labeled only toxic, but not hateful or offensive.	50	6.08%
HO	Comments labeled both hateful and offensive, but not toxic.	13	1.58%
OT	Comments labeled both offensive and toxic, but not hateful.	117	14.22%
HT	Comments labeled both hateful and toxic, but not offensive.	15	1.82%
HOT	Comments labeled as all three of hateful, offensive, and toxic.	99	12.03%
Overlap (any)	Comments labeled as any of hateful, offensive, and toxic.	244	29.65%

Table Notes: The leftmost column indicates binary outcome variable names. The center column defines each outcome variable. The two rightmost columns indicate how many and the percentage of comments out of 823 total directed comments received a TRUE value.

For regression analyses, we used 11 binary outcome variables (see Table 2 for overview of all outcome variables). Analyses of H (all), O (all), and T (all) outcome variables helped us understand features of an *entire* concept. Analyses of H (only), O (only), and T (only) outcome variables helped us understand features unique to comments labeled as *only one* concept. Subsets featuring overlapping concepts as outcome variables – HO, HT, OT, HOT, or any overlap – helped us identify how comment features intersect between concepts. When including all variables as predictors, most reported models (see Results) offered a statistically significant better fit than null models at a .05 level as per likelihood ratio tests (see Appendix Table 7 for likelihood ratio test results) except for H (Only), O (Only), HO, HT, which can be expected given the small number of comments with a value of TRUE for each of these models (see Table 2 for count of TRUE values for each outcome variable). In order to understand meaningful differences between H (Only), O (Only), HO, HT, and other concepts through statistically significant models, we performed variable selection based on theory and observed variable significance for these outcome variables. For example, given the relatively small number of comments included in the H (only) sample, we elected to drop all insignificant variables for H (only) in favor of increasing degrees of freedom and providing meaningful insight about the single observed significant variable in our original model – *sarcasm*. For HT, while within *pronoun use* only '2nd or 3rd Person' was statistically significant, we still included '1st Person' since we could not identify a reasonable theoretical basis on which to exclude it while still including '2nd or 3rd Person'. The resulting models for binary outcome variables H (Only), O (Only), HO, and HT thus included a smaller subset of predictors than all other outcome variables. This variable selection means that we are unable to make certain comparisons between concepts involving H (Only), O (Only), HO, and HT.

4 RESULTS

Our analyses find that our expectations are weakly met for **H1**, strongly met for **H2** and **H3**, and ambiguous results for **H4**. Specifically, we find that O experiences the highest interannotator agreement (**H1**), *directedness* at groups predicts H (**H2**), *aggressiveness* predicts O and *sarcasm* and

threats predict T (**H3**), and *rater effect* predicts all concepts (**H4**). Subsections will refer to Table 3, which represents regression model outputs for each of our eleven outcome variables.

4.1 O Experiences Highest Interannotator Agreement

H1: H and O concepts will experience more consistent annotations across annotators than T.

Our calculations of Krippendorff's alpha between annotators (see Figure 2) indicate that our expectations were met, albeit weakly. As expected, we see that O and H experience the highest measures of agreement of the three concepts. However, the 95% confidence interval for O overlaps with parts of both H and T's confidence intervals. H's confidence interval entirely encompasses T's and partially encompasses O's. In this sense, the statistical significance of our findings indicates that our expectations were only weakly met.

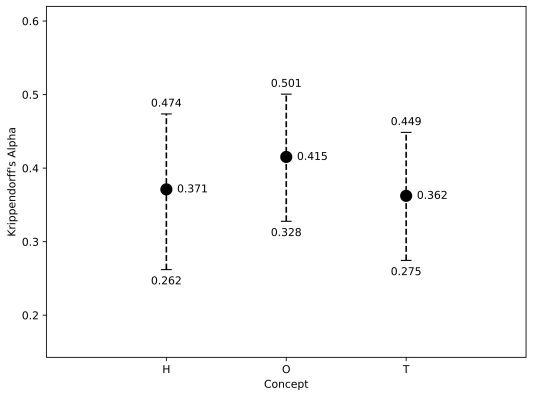

Fig. 2. Krippendorff's alpha scores for each HOT concept with 95% confidence intervals given as the 2.5th and 97.5th percentiles of the bootstrap distribution.

Table 3. Statistically significant predictors for directed comments in odds ratios.

	Dependent variable:										
	H (all)	O (all)	T (all)	H (only)	O (only)	T (only)	HO	HT	OT	HOT	Overlap (any)
Communication Style											
Assertive	0.857***	0.736***	0.820***		0.946**	1.051**			0.891***	0.885***	0.768***
Passive	0.820**	0.691***	0.957		0.927	1.342***			0.857*	0.858*	0.713***
Passive-Aggressive	0.900***	0.776***	0.842***		0.987	1.066***			0.865***	0.918***	0.783***
Sarcasm	0.975	0.928	1.188***	0.972	0.940*	1.255***			0.973	1.004	0.961
Pronoun Use											
1st Person	0.884	1.007	0.788			0.976		0.912	0.962	0.943	0.833
2nd or 3rd Person	0.944	1.046	0.999			1.042		0.913**	1.045	1.002	0.970
Directed at											
Group or Individual	1.335*	1.248	1.530**			1.133			0.989	1.374**	1.333
Behavior	1.229	1.205	1.392*			1.099			0.981	1.339**	1.243
Behavior											
Accusation	1.181***	1.004	1.084*			1.014			0.922**	1.137***	1.072*
Benign	1.197	1.025	0.993			0.883*			0.957	1.167	1.139
Threat	1.154**	1.120	1.411***			1.171***			1.041	1.142**	1.195**
Insult	1.137***	1.127**	1.151***			0.999			1.019	1.118***	1.158***
Misinformation	1.129	1.069	1.032			0.949			0.985	1.072	1.063
Name-calling	1.277***	1.313***	1.368***			1.013	1.021**	1.026*	1.093**	1.198***	1.376***
Rater Effect											
H	5.750***									2.633***	1.730
O		4.690***							1.922***	1.073	2.249**
T			7.298***			1.709***		1.141	1.274	1.518*	2.076**
Constant	0.626**	0.699	0.474***	1.028***	1.094***	0.737**	1.009*	1.058	0.927	0.551***	0.581**
Observations	823	823	823	823	823	823	823	823	823	823	823
Log Likelihood	−319.061	−442.662	−448.148	355.090	−45.276	91.732	549.717	496.403	−251.955	−192.227	−401.520
Akaike Inf. Crit.	670.122	917.323	928.296	−678.179	122.552	−151.465	−1,065.435	−958.805	537.909	420.454	839.041

Table Notes: *p<0.1; **p<0.05; ***p<0.01

For binary variable "sarcasm", the holdout category is "n" or no sarcasm present. For categorical variable "communication style", the holdout category is "aggressive". In order to understand meaningful differences between H (Only), O (Only), HO, HT, and other concepts through statistically significant models, we performed variable selection based on theory and observed variable significance for these outcome variables. The resulting models for binary outcome variables H (Only), O (Only), HO, and HT thus included a smaller subset of predictors than all other outcome variables, demonstrated by the empty values for some variables under these models.

4.2 Directedness at Groups Predicts H, Aggressiveness Predicts O, Sarcasm and Threats Predict T

H2: *Directedness* at groups will be a predictive feature of comments labeled as H.
H3: All comments will be predicted by comment features outside of those made explicit in concept definitions.

For **H2**, our expectations were met that *directedness* at groups is a statistically significant predictor of all comments labeled with H (OR=1.335, see Venn diagram in Figure 3 for additional context).

H is more likely to be *directed* at groups or individuals.

However, *directedness* at groups was not H's only predictive feature. Like for **H2**, for **H3** our expectations were met that all comment labels, including H, are predicted by various comment features not made explicit in concept definitions (see Figure 3 for Venn diagram characterizing positively predictive features for each outcome variable). When looking at the subsets of H, O, and T that include comments labeled with *only* H, O, or T, we find no significant predictors of comments labeled with only H, owing likely to this subset's small sample size. Comments labeled with only O are less likely to be *assertive* than *aggressive* (OR=0.946), and less likely to be *sarcastic* (OR=0.940). Comments labeled with only T are more likely to be *assertive* (OR=1.051), *passive* (OR=1.342), and *passive-aggressive* (OR=1.066) than *aggressive*; *sarcastic* (OR=1.255); and feature *threats* (OR=1.171, see Figure 3), and are *less* likely to be *benign* (OR=0.883).

O is more likely to be *aggressive* and less likely to be *sarcastic*; T is less likely to be *aggressive* and more likely to be *sarcastic* and feature *threats*.

When looking at *all* comments labeled with H, O, and T, we see similar patterns, but concepts appear to be less distinct from one another. For example, for all comments labeled with T, like comments labeled *only* with T, being *sarcastic* (OR=1.188) and featuring *threats* (OR=1.411) are significant predictors. In addition, being *directed at* behaviors (OR=1.391) and groups or individuals (OR=1.530), and featuring *accusations* are significant predictors (OR=1.084). Directly opposite to comments labeled with *only* T, *all* comments labeled with T are less likely to be *assertive* (OR=0.820) and *passive-aggressive* (OR=0.842) than *aggressive*. These predictors partially overlap with significant predictors of all comments labeled with H, which as per **H2** include *directedness at* groups or individuals, as well as featuring *accusations* or *threats* and being less likely to be *assertive*, *passive*, and *passive-aggressive* than *aggressive*. Finally, all comments labeled feature the same significant predictors as comments labeled with *only* O, and are additionally less likely to be *passive* (H (all) OR=0.820, O (all) OR=0.691, T (all) OR=0.957) than *aggressive*. As a whole, all comments labeled with concepts are less distinct from other concepts than comments labeled with only a particular concept. This finding is especially observable in that all three concepts are more likely to include *insult* and *name-calling* and be affected by raters' biases (i.e. *rater effect* predictor – H (all) OR=5.750, O (all) OR=4.690, T (all) OR=7.298).

H, O, and T are all likely to include *insults* and *name-calling* and to be predicted by *rater effect*.

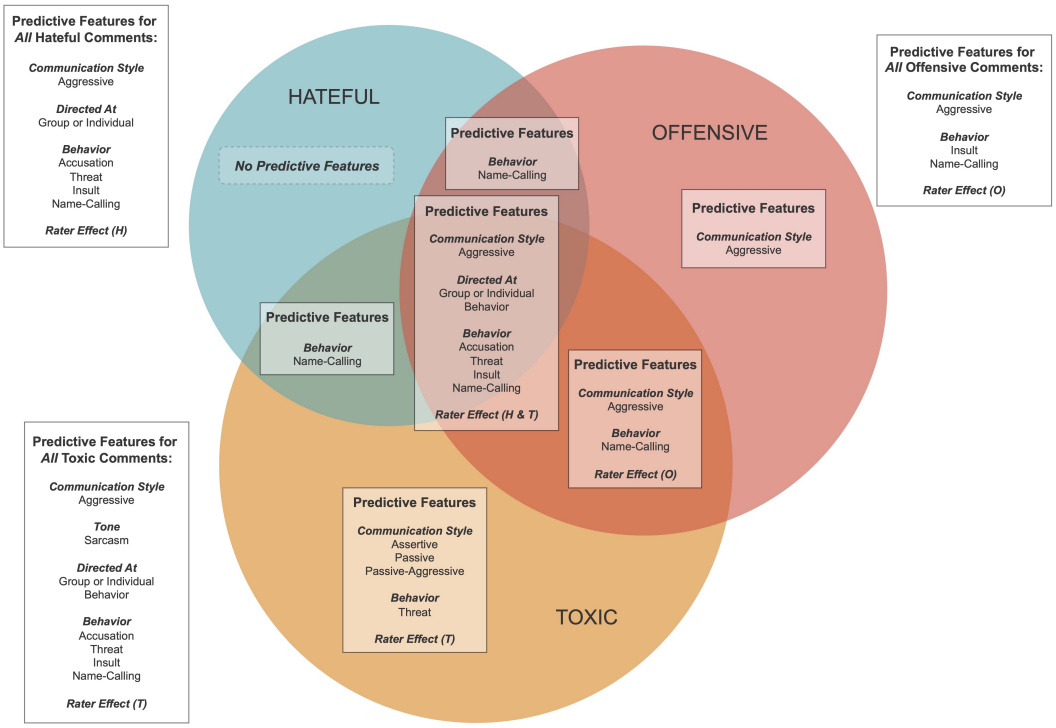

Fig. 3. Venn diagram of positive predictors of each concept subset based on regression analysis and odds ratios in Table 3. Note that this figure does *not* include negatively predictive variables.

4.3 Rater Effect Predicts All Concepts

H4: Individual raters' biases will be more predictive of T label outcomes than either of H and O label outcomes.

Results concerning **H4** are ambiguous. Specifically, in model results with outcome variables referring to all comments labeled with H, O, or T, *rater effects* are statistically significant and exhibit the greatest magnitude among predictors for H, O, and T. However, for comments featuring any overlapping concepts, O and T *rater effects* are significant. For OT comments, O *rater effect* is significant. For HOT comments, H and T *rater effects* are significant. These subset analyses suggest that *rater effect* is most commonly a significant predictor for T in subsets, followed by O then H. Since *rater effect* is the most commonly a significant predictor for T relative to the other two concepts, our expectations from **H4** are partially met. However, whether statistically significant or not, *rater effect* has the highest magnitude effect on all outcome variables. In this sense, results concerning **H4** are ambiguous.

5 DISCUSSION

We evaluated assumptions about off-the-shelf (OTS) models reusability for detecting various types of harmful content and identified comment features, both specified in harm definitions and not, that predict how models will label data. The features we identified include: *communication style*; *sarcasm*; who comments are *directed* at; insults, name-calling, and threats; and annotators' subjective interpretive lenses. These findings show that model reusers cannot assume that a model designed

to detect a particular concept will truly detect that concept. Instead, model reusers can turn to concept definitions, annotation task design, and additional features specified in our codebook to make sense of expected model output to determine whether that output aligns with their goals for a new harmful content detection task like moderating an online community or website. Based on our findings, we offer a decision tree for how model reusers can assess an OTS model's fit for a reuser's needs in Figure 4, including examples of how to apply our codebook to concepts beyond hateful, offensive, and toxic in Table 4,

The following subsections provide a user guide for this decision tree, discussing each numbered box of the decision tree in detail. We conclude by discussing additional findings that future work should investigate.

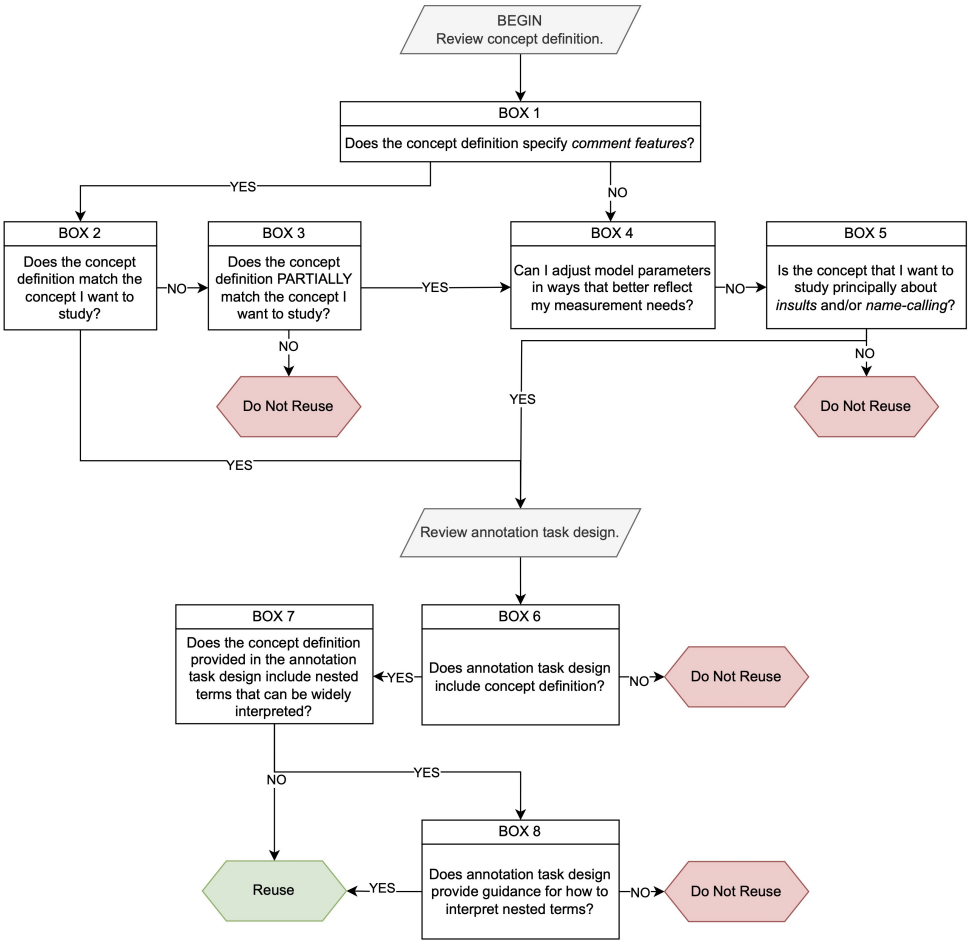

Fig. 4. OTS model reuse decision tree.

5.1 [BOX 1] Does the concept definition specify *comment features*?

The more a harm concept definition specifies *comment features* in particular (i.e., O's definition), the more consistently annotators may interpret annotation instructions and assign labels to comments.

This finding seconds Pustejovsky and Stubbs' [42] recommendation that the more specific a concept definition, the more consistent labels will be across annotators. Similarly, Li et al.'s [37] find that when ChatGPT in an annotator role was asked to respond to various annotation tasks (i.e., prompts) providing more or less detail about how to identify H, O, and T, the more detail they provided ChatGPT, the more reliable and consistent ChatGPT's annotations were. If an OTS model's harm concept definition and annotation task design (see Box 6-8 for more detail on reviewing annotation task design) specify comment features that annotators can use to identify the presence of that concept, model reusers can reasonably expect the model to detect that concept.

To assess whether the concept definition specifies comment features, we suggest applying our comment feature codebook to the harm concept of interest. Table 4 provides examples of applying our codebook to harm concept definitions beyond H, O, and T. We used *incivility* and *impoliteness* as defined by Papacharissi [39] as example concepts that political communications researchers have studied [45, 46]. As Table 4 shows, by applying our comment feature codebook, it is possible to identify both comment features specified in definitions of incivility and impoliteness, as well as nested terms (see Boxes 7-8 for more on nested terms). If the harm concept definition specifies comment features, a model reuser can move on to Box 2. If not, a model reuser can move on to Box 4.

Table 4. Examples of two harm concepts – incivility and impoliteness – annotated with our explanatory features codebook.

Concept	Definition Annotated with Explanatory Features Codebook
Incivility	"(1) Does the discussant verbalize a threat [**threat**] to democracy [**directedness** toward a **group**](e.g.propose to overthrow a democratic government by force)? (2) Does the discussant assign stereotypes (e.g. associate a person with a group by using labels, whether those are mild – 'liberal', or more offensive – 'faggot') [**name-calling**]? (3) Does the discussant threaten other individuals' rights (e.g. personal freedom, freedom to speak) [**aggressive** or **passive-aggressive**]." [39]
Impoliteness	"If name-calling (e.g. weirdo, traitor, crackpot) [**name-calling**], aspersions (e.g. reckless, irrational, un-American) [**insult** or **accusation**], synonyms for liar (e.g. hoax, farce) [**name-calling**], hyperbole (e.g. outrageous, heinous) [**could be added as a new dimension to codebook**], words that indicated non-cooperation [**nested term that requires definition**], pejorative speak [**insult**], or vulgarity [**nested term that requires definition**] occurred, then the message was considered impolite." [39]

5.2 [BOX 2] Does the concept definition match the concept I want to study?

In this step, if a model reuser identifies that the OTS model's concept definition matches their concept of interest, our findings suggest that the OTS model would be an appropriate candidate for reuse, and the reuser can move onto Box 6. If the definition does not match the reuser's concept of interest, the reuser can move on to Box 3. This step may seem simple, but as we discussed in our introduction, modelers have not always been explicit about their concept definitions, and as such it is important for reusers to investigate the models' definitions and assumptions against reusers own goals.

5.3 [BOX 3] Does the concept definition PARTIALLY match the concept I want to study?

In this step, if a model resuser determines that the OTS model's concept definition only partially matches their concept of interest, the reuser can move on to Box 4. If the definition and specifications

do not even partially match the reuser's concept of interest, the OTS model is not an appropriate candidate for reuse, and reusers should look for a different model.

5.4 [BOX 4] Can I adjust model parameters in ways that better reflect my measurement needs?

If an OTS model's harm concept definition does not specify comment features that annotators can use to identify the presence of that concept, our findings suggest that the model reuser cannot expect the model to consistently detect the concept. In this case, we encourage reusers to investigate whether they can adapt the OTS model to better reflect the reuser's concept of interest. For example, in addition to toxicity scores for content, Jigsaw's Perspective API returns attributes 'insult', 'identity attack', 'profanity', 'threat', and 'severe toxicity' as per their definitions of these terms.[5] While it is not possible for a reuser external to Perspective API's development team to change the weights of these attributes when using Perspective API to label new datasets, a reuser could consider developing their own model trained on a content sample labeled by Perspective API, but with differently weighted attributes. Using this approach, a reuser could cobble together a weighted definition of harm reliant on more specific comment features that more closely matches their reuse needs. If a reuser takes a parameter adjustment approach, we recommend that the reuser provide detailed documentation about their adaptations and modifications, and to check for nested terms in all adapted parameters as per Box 8.

In this vein, we encourage well-resourced institutions like Jigsaw and OpenAI that generously make their harmful content detection algorithms available for public use to consider how to make their resources adaptable to specific harmful content detection needs. For example, these organizations could solicit requests from their API users about which additional harm-related attributes would be helpful so that reusers can more easily construct bespoke models with attributes that match their harm-detection needs. Just as validated survey instruments can be adapted to some degree for contextually-specific research goals, we encourage these organizations to make their harmful content detection models adaptable for contextually-specific harmful content detection moderation and research goals. Alternatively or in addition, these institutions can consider providing access to their encoders or classifiers to support model reusers in applying existing techniques for adapting models trained to detect similar, but not identical, concepts [29].

If it is possible for a reuser to adapt an OTS model in ways that more closely reflect their needs, the model may be an appropriate candidate for reuse and the reuser can move onto Box 6. If not, the reuser can move on to Box 5.

5.5 [BOX 5] Is the concept that I want to study principally about *insults* and/or *name-calling*?

All three of H, O, and T concepts analyzed share the common predictive features of *insults* and *name-calling*. It is possible that these common features reflect how annotators' perceive harm in general, and since each of H, O, and T refer to variations of harm, annotators tended to look for insults and name-calling as identifying features of each concept even though Wu et al. [60] did not specifically ask them to. In order to assess whether this explanation is accurate, we encourage future research to study how annotators perceive and define 'harm' in general when they review social media comments when they are not provided with a specific definition of harm in advance.

Given these findings, if a reuser's concept of interest differs from the concept a reusable model is trained to detect, but is mostly about *insults* or *name-calling*, we show that using any of these three

[5]See https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US for definitions of these attributes.

OTS models – and perhaps others beyond these models – may still be an appropriate candidate for reuse provided sufficient contextualization.

Contextualization here is about clarifying the limits of inferences and may include identifying which additional definitional features reusers are interested in studying that are *not* captured by OTS models, clearly describing the data to which reusers are applying OTS models, and articulating specific limitations associated with reusing the OTS model. Contextualizing model reuse can help to ensure that findings are not inappropriately generalized or ineffectively used to address a potentially unrelated harmful content detection issue.

If the reuser is principally interested in understanding *insults* or *name-calling*, they can move on to Box 6. If the reuser is not principally interested in identifying or studying *insults* and/or *name-calling*, we do not recommend reusing the OTS model.

5.6 [BOX 6] Does annotation task design include concept definition?

One might assume that an annotation task to detect a particular harm concept would include the harm concept's definition, but this is not always the case. For example, while Davidson et al.'s [10] model was trained on a dataset for which annotators were asked to distinguish between hateful and offensive content, the annotation task design for this task did not include a definition of offensiveness. Given lack of specificity in defining offensiveness, it is difficult to assess whether Davidson et al.'s model is actually trained to detect offensiveness. In a related example, Aluru et al.'s [2] H model is trained on several datasets that define H differently, making it difficult to assess which definition of hatefulness Aluru et al.'s model detects. In these instances, we recommend considering an alternative model that provides the harm concept's definition in its annotation task design.

In other instances, concept definitions may be clearly included as part of the annotation task design, as in Stoll et al.'s [53] incivility and impoliteness prediction models for German language texts. If the annotation task design includes the harm concept's definition, the reuser can move on to Box 7

5.7 [BOX 7] Does the concept definition include nested terms that can be widely interpreted?

Beyond *comment features*, definitions of H, O, and T that Wu et al. used specified *author features* and *audience features* (i.e., *intent* and *effect* respectively) that annotators could look for to identify the presence of each concept. However, Wu et al. did not provide annotators with information about author *intent* and audience *effect* with which to make labeling decisions. *Author features* and *audience features* are types of what we call nested terms that we introduced in relation to **H3**, or terms that are embedded in another term's concept definition. For example, T's definition referred to an *effect* on an audience – "likely to make readers want to leave a discussion" – but annotators were not given information about whether audiences actually wanted to leave discussions. It is possible that in the absence of this additional information, the predictive feature *threat* describes how annotators interpreted and applied this anticipated *effect*. Similarly, *sarcastic* may reflect annotators' interpretation and application of the nested term "a rude, disrespectful, or unreasonable"'s *effect* on an audience specified in T's definition. As such, these predictive features may represent annotators' interpretations of nested terms. However, since Wu et al. did not ask annotators specifically about how they interpreted and applied nested terms included in T's definition, it is possible that this is not the case. Future research should assess whether it is possible to use *behaviors* and *communication styles* as proxies for nested terms, including *author features* and *audience features*.

In the absence of this additional research, if a harm concept definition provided in an annotation task includes nested terms, the model reuser can move on to Box 8. If the definition does not include nested terms, our findings suggest that the OTS model would be an appropriate candidate for reuse.

5.8 [BOX 8] Does the annotation task design provide guidance for how to interpret nested terms?

In the event that an annotation task's term definition includes nested terms, consider whether the task design also provides guidance on how to interpret these nested terms. For example, in the case of Jigsaw's toxicity, a reuser could investigate how Jigsaw instructed annotators to assess the nested term: "likely to make readers want to leave a discussion". Did Jigsaw specify that "leave a discussion" means that no users commented after the comment in question? How did Jigsaw ask annotators to assess comments for the presence of this nested term? Were comment threads from before and after the comment in question provided to assess the presence of this nested term? If this information is difficult to find, model reusers can investigate whether evidence exists about how annotators interpreted nested terms. Returning to the toxicity example, a study by Saveski et al. finds that Tweets labeled as toxic by Perspective API actually tend to generate *more* responses and deeper reply trees than non-toxic Tweets [47]. This finding suggests that Jigsaw's inclusion of the "leave a discussion" nested term may have had little bearing on how annotators labeling training data identified toxic content and, consequently, what Perspective API identifies as toxic content. While it is possible that Jigsaw provided detail about how to operationalize "leave a discussion" in their annotation task design, Saveski et al.'s study indicates that whether they did or not, "leave a discussion" does not seem to be a characteristic feature of how Perspective API operationalized toxicity, suggesting that whatever detail Jigsaw provided was likely insufficient for annotators to interpret consistently.

When nested terms occur, we thus encourage model reusers to consider whether the annotation task provides sufficient guidance for how to interpret nested terms consistently and if not, whether evidence exists about how annotators have interpreted nested terms. If so, the OTS model may be an appropriate candidate for reuse. If the annotation task does not provide sufficient guidance about how to interpret nested terms and insufficient evidence exists about how annotators interpreted nested terms in practice, we encourage reusers to consider an alternative OTS model.

5.9 Additional Findings & Directions for Future Work

5.9.1 Rater Effect. Beyond comment and definition features, our findings show that *rater effect* has the highest magnitude effect size for all of H, O, and T. That *rater effect* has the highest significant effect size across H, O, and T concepts opens important opportunities for future work to investigate *how* individual annotators' perspectives – and consequently groups of annotators' perspectives – affect training data. Would Wu et al.'s dataset have received similar labels if their annotator pool were majority gender non-binary rather than majority male or female, for example? What if annotators were labeling under different labor conditions – fixed salary and enforced breaks, for instance? We encourage future research to investigate questions about annotators' demographics and experiences that likely influence their interpretations and labels.

Despite having the highest magnitude effect size across harm concepts, as expected hypothesis **H4**, *rater effect* is most commonly a significant predictor for T. It is possible that *rater effect* is significant for H less frequently across outcome variables than for O and T because H's definition provides information to annotators about specific *comment features* to identify. Conversely, it is possible that *rater effect* is most often significant for T relative to the other two concepts because T's definition refers only to author *intent* or the *effect* of a comment on its audience. Annotators did not have direct information about either of these features during the annotation task, potentially

making individual raters' interpretive differences more impactful. This explanation for some of our findings echoes Pustejovsky and Stubbs' [42] recommendation that the more specific a concept definition is the less space there is for differences in interpretation to affect annotators' labeling decisions.

5.9.2 A Word of Caution to Social Science Researchers. Our results show that Krippendorff's alphas between annotators remain below 0.5. Conventionally, social science research determines that reliable conclusions about social phenomena derived from an annotated dataset can be made when Krippendorff's alpha is greater than or equal to 0.8, with an alpha of 0.667 being the lower limit for making tentative conclusions from a dataset [33]. Given that OTS models are often reused in social science research to study harm in online communities, the degree of human disagreement in this replication attempt of OTS models' training datasets indicates that social science researchers should exercise caution when using OTS models to study and make conclusions about social phenomena.

5.9.3 Validating OTS Model Reuse Decision Tree. We encourage future research that validates the usefulness and feasibility of implementing the OTS Model Reuse Decision Tree in practice. For example, researchers could conduct user studies with model reusers applying the Decision Tree to their own OTS model reuse tasks, assessing the Decision Tree's usefulness relative to their tasks. Future work could iterate on the Decision Tree based on user feedback. Future validation effort could also include reaching out to developers of popular OTS models to assess the feasibility of enabling user adaptability and modification, and refine recommendations for OTS model creators based on developers' feedback and insights. Collectively, these future steps will help to ensure that our OTS Model Reuse Decision Tree is as useful and feasible as possible.

6 CONCLUSION

Our study develops best practices for evaluating whether and how to reuse an OTS harmful content detection model. By using content analysis and statistical methods to evaluate assumptions about OTS model utility and reusability, we show that people attempting to reuse a model cannot assume that a model apparently designed to detect a particular concept, will actually detect that concept. Instead, our proposed best practices for OTS model reuse direct reusers to assess concept definitions, annotation task design, and additional features specified in our codebook to identify expected model output, and consequently, determine whether that OTS model is appropriate for a particular reuse. We also recommend strategies for model reusers to adapt OTS models when possible and contextualize OTS model reuse when appropriate, and for OTS model providers to consider how to make their models more adaptable to contextually-specific harmful content detection goals. We contribute to the broader discussion on responsible, ethical AI use [31] that encourages researchers to ensure that models are fit-for-purpose [6] by providing guidance for ensuring model validity and identifying biases. Automated harmful content detection persistently navigates complex tensions between supporting online safety while also encouraging social inclusion [19]. In this challenging work, model validity and identifying biases remain pressing concerns for automated harmful content detection with consequences of mis-, over-, or under-identifying harmful content affecting inclusive access to digital resources [38] and societal wellbeing on- and offline [61]. Our research shows that while OTS model reuse can be an important strategy in resource-constrained environments and to mitigate environmental harms associated with training new models, reusers must critically and deliberately attend to those OTS models' specifications and when considering them for reuse for new harmful content detection tasks.

ACKNOWLEDGMENTS

We are thankful for our colleagues' help and feedback on this project. Paul Resnick provided comments on early analyses. Morgan Wofford provided comments on our content analysis process. Corey Powell and Abner Heredia at Consulting for Statistics, Computing, and Analytics Research (CSCAR) at the University of Michigan advised on statistical analysis. Shubham Atreja, Gina Brandolino, Lizhou Fan, Allegra Fonda-Bonardi, Ji Eun Kim, Lingyao Li, and Allie Piippo provided comments on earlier drafts. Nico Wilkins provided copyediting services.

REFERENCES

[1] 2022. Sarcasm. In *Cambridge Advanced Learner's Dictionary & Thesaurus*. Cambridge University Press.
[2] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. A Deep Dive into Multilingual Hate Speech Classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V* (Ghent, Belgium). Springer-Verlag, Berlin, Heidelberg, 423–439.
[3] Sai Saketh Aluru, Punyajoy Saha, and Binny Mathew. 2020. DE-LIMIT: DeEpLearning models for MultIlingual haTespeech (DELIMIT): Benchmarking multilingual models across 9 languages and 16 datasets.
[4] Amazon Web Services. 2016. Amazon Rekognition.
[5] Perspectiv API. [n. d.]. Case Studies. https://perspectiveapi.com/case-studies/. Accessed: 2024-6-26.
[6] Jacqui Ayling and Adriane Chapman. 2022. Putting AI ethics to work: are the tools fit for purpose? *AI Ethics* 2, 3 (Aug. 2022), 405–429.
[7] Nicole A Beres, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21, Article 438)*. Association for Computing Machinery, New York, NY, USA, 1–15.
[8] Matthew Blackwell, James Honaker, and Gary King. 2017. A unified approach to measurement error and missing data: Overview and applications. *Sociol. Methods Res.* 46, 3 (Aug. 2017), 303–341.
[9] Kevin Coe, Kate Kenski, and Stephen A Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *J. Commun.* 64, 4 (Aug. 2014), 658–679.
[10] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.
[11] Thomas Davidson, Ingmar Weber, and Jonathan Zarecki. 2019. Automated Hate Speech Detection and the Problem of Offensive Language.
[12] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 67–73. https://doi.org/10.1145/3278721.3278729
[13] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media*.
[14] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Twelfth International AAAI Conference on Web and Social Media*.
[15] André Ferreira. 2022. Analyzing Political Manifestos through Natural Language Processing and Dashboarding. https://towardsdatascience.com/analyzing-political-manifestos-through-natural-language-processing-and-dashboarding-4ad1d62d6b9a. Accessed: 2024-6-26.
[16] André Ferreira. 2022. Polids. https://andrecnf-polids-appapp-naawtf.streamlit.app/. Accessed: 2024-6-26.
[17] Christian Fong and Matthew Tyler. 2021. Machine Learning Predictions as Regression Covariates. *Political Analysis* 29, 4 (2021), 467–484. https://doi.org/10.1017/pan.2020.38
[18] Loshini Ganeshan, Masitah Ghazali, and Nur Zuraifah Syazrah Othman. 2023. An Acceptance Towards Buzzer as a Filtering Approach towards Creating Responsible Users on Social Media Postings. In *Proceedings of the Asian HCI Symposium 2022* (New Orleans, LA, USA) *(Asian HCI '22)*. Association for Computing Machinery, New York, NY, USA, 33–39.
[19] Rosalie Gillett, Zahra Stardust, and Jean Burgess. 2022. Safety for whom? Investigating how platforms frame and perform safety and harm interventions. *Soc. Media Soc.* 8, 4 (Oct. 2022), 205630512211443.
[20] Tesh Goyal, Ian Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Jigsaw Specialized Rater Pools Dataset.

[21] 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* (Toronto, Canada) *(AISec '18)*. Association for Computing Machinery, New York, NY, USA, 2–12.
[22] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv [cs.CL]* (April 2020).
[23] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *AI Open* 2 (Jan. 2021), 225–250.
[24] Hatebase. [n. d.]. Hatebase. https://hatebase.org/. Accessed: 2022-8-4.
[25] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qual. Health Res.* 15, 9 (Nov. 2005), 1277–1288.
[26] Kokil Jaidka, Subhayan Mukerjee, and Yphtach Lelkes. 2023. Silenced on social media: the gatekeeping functions of shadowbans in the American Twitterverse. *J. Commun.* 73, 2 (April 2023), 163–178.
[27] Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean Offensive Language Dataset. (Dec. 2022), 10818–10833.
[28] Jigsaw. 2017. Perspective API. www.perspectiveapi.com. Accessed: 2022-1-NA.
[29] Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Stroudsburg, PA, USA, 3770–3783.
[30] Mateusz Kazimierczak, Thanyathorn Thanapattheerakul, and Jonathan H Chan. 2023. Enhancing Security in WhatsApp: A System for Detecting Malicious and Inappropriate Content. In *Proceedings of the 12th International Symposium on Information and Communication Technology* (<conf-loc>, <city>Ho Chi Minh</city>, <country>Vietnam</country>, </conf-loc>) *(SOICT '23)*. Association for Computing Machinery, New York, NY, USA, 274–281.
[31] Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *J. Artif. Intell. Res.* 71 (July 2021), 431–478.
[32] Klaus Krippendorff. 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educ. Psychol. Meas.* 30, 1 (April 1970), 61–70.
[33] Klaus Krippendorff. 2006. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research* 30, 3 (01 2006), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x arXiv:https://academic.oup.com/hcr/article-pdf/30/3/411/22338169/jhumcom0411.pdf
[34] Klaus Krippendorff. 2019. Conceptual Foundation. In *Content Analysis: An Introduction to Its Methodology* (4 ed.), Klaus Krippendorff (Ed.). SAGE, 24–50.
[35] Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to Avoid: Demoting Latent Confounds in Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4153–4163. https://doi.org/10.18653/v1/D19-1425
[36] Tu Le, Danny Yuxing Huang, Noah Apthorpe, and Yuan Tian. 2022. SkillBot: Identifying Risky Content for Children in Alexa Skills. *ACM Trans. Internet Technol.* 22, 3 (July 2022), 1–31.
[37] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "HOT" ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media. *ACM Trans. Web* 18, 2, Article 30 (mar 2024), 36 pages. https://doi.org/10.1145/3643829
[38] Samuel Mayworm, Shannon Li, Hibby Thach, Daniel Delmonaco, Christian Paneda, Andrea Wegner, and Oliver L Haimson. 2024. The Online Identity Help Center: Designing and developing a content moderation policy resource for marginalized social media users. *Proc. ACM Hum. Comput. Interact.* 8, CSCW1 (April 2024), 1–30.
[39] Zizi Papacharissi. 2004. Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society* 6, 2 (2004), 259–283. https://doi.org/10.1177/1461444804041444 arXiv:https://doi.org/10.1177/1461444804041444
[40] Dark Data Project. 2023. About. https://darkdataproject.org/about. Accessed: 2023-8-8.
[41] The Sentinel Project. 2018. Home. https://thesentinelproject.org/. Accessed: 2023-8-8.
[42] James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc.
[43] R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
[44] Sarah T Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

[45] Patrícia Rossini. 2022. Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk. *Communication Research* 49, 3 (2022), 399–425. https://doi.org/10.1177/0093650220921314 arXiv:https://doi.org/10.1177/0093650220921314

[46] Ian Rowe. 2015. Civility 2.0: a comparative analysis of incivility in online political discussion. *Information, Communication & Society* 18, 2 (2015), 121–138. https://doi.org/10.1080/1369118X.2014.940365 arXiv:https://doi.org/10.1080/1369118X.2014.940365

[47] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1086–1097. https://doi.org/10.1145/3442381.3449861

[48] Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. 2022. Reduce, Reuse, Recycle: Green Information Retrieval Research. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Madrid</city>, <country>Spain</country>, </conf-loc>) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2825–2837.

[49] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International workshop on natural language processing for social media*. 1–10.

[50] Angela M Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2022. Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors. *New Media & Society* (Dec. 2022), 14614448221138529.

[51] Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive Language and Hate Speech Detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 3498–3508.

[52] Spandana Singh. 2019. *Everything in Moderation: The Limitations of Automated Tools in Content Moderation.* Technical Report 1. Open Technology Institute.

[53] Anke Stoll, Marc Ziegele, and Oliver Quiring. 2020. Detecting Impoliteness and Incivility in Online Discussions. *Computational Communication Research* 2, 1 (2020), 109–134. https://doi.org/10.5117/CCR2020.1.005.KATH

[54] Nathan TeBlunthuis, Valerie Hase, and Chung-Hong Chan. 2024. Misclassification in automated content analysis causes bias in regression. Can we fix it? Yes we can! *Commun. Methods Meas.* 18, 3 (July 2024), 278–299.

[55] University of Kentucky Violence Intervention and Prevention Center 2014. *The Four Basic Styles of Communication.* University of Kentucky Violence Intervention and Prevention Center.

[56] Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS One* 15, 12 (Dec. 2020), e0243300.

[57] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada, 78–84.

[58] 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.

[59] M Wiegand, M Siegel, and J Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of the GermEval 2018 Workshop*.

[60] Siqi Wu, Angela Schöpke-Gonzalez, Sagar Kumar, Libby Hemphill, and Paul Resnick. 2023. HOT Speech: Comments from Political News Posts and Videos that were Annotated for Hateful, Offensive, and Toxic Content.

[61] Greyson K Young. 2022. How much is too much: the difficulties of social media content moderation. *Inf. Commun. Technol. Law* 31, 1 (Jan. 2022), 1–16.

[62] Savvas Zannettou, Mai Elsherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and Characterizing Hate Speech on News Websites. In *Proceedings of the 12th ACM Conference on Web Science* (Southampton, United Kingdom) *(WebSci '20)*. Association for Computing Machinery, New York, NY, USA, 125–134.

[63] Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. 2016. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? *BMC Med. Res. Methodol.* 16 (Aug. 2016), 93.

A HARM CONCEPT REVIEW

Table 5. Review of Harm Concept Definitions and Features

Term	Definition	Definitional Features
Hate	"...language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group" (Davidson et al., 2017, p. 1)	Directedness, who or what a comment is directed at, author intent
Hate	Language that: "...uses a sexist or racial slur. attacks a minority. seeks to silence a minority. criticizes a minority (without a well founded argument). promotes, but does not directly use, hate speech or violent crime. criticizes a minority and uses a straw man argument. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims. shows support of problematic hashtags. E.g. "#BanIslam", "#whoriental", "#whitegenocide" negatively stereotypes a minority. defends xenophobia or sexism. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria." (Waseem & Hovy, 2016, p. 89)	Directedness, who or what a comment is directed at, author intent, type of behavior exhibited by author
Hate	"...denigrates a person because of their innate and protected characteristics" (ElSherief, Kulkarni, et al., 2018, p. 1; ElSherief, Nilizadeh, et al., 2018, p. 52)	Directedness, who or what a comment is directed at, effect of a comment on readers
Hate	"...any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic" Nockleby by way of Schmidt and Wiegand (2017, p. 1)	Directedness, who or what a comment is directed at, effect of a comment on readers
Offensive	Specific words, phrases, or collections of symbols that can collectively be interpreted as hateful (Hatebase, n.d.)	Presence of specific terms
Offensive	"hurtful, derogatory or obscene comments made by one person to another person" (Wiegand et al., 2018, p. 1)	Directedness, who or what a comment is directed at, effect of a comment on readers, presence of specific terms
Toxic	"a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion" (Perspective API via Wulczyn et al. 2017)	Effect of a comment on readers
Toxic	"various types of negative behaviors involving abusive communications directed towards other players (i.e., harassment, verbal abuse, and flaming) and disruptive gameplay that violates the rules and social norms of the game (i.e., griefing, spamming, and cheating)" (Beres et al., 2021, p. 1)	Directedness, type of behavior exhibited by author, who or what a comment is directed at, effect of a comment on readers
Toxic	"the use of profane language by one player to insult or humiliate a different player in his own team" (Märtens et al., 2015)	Directedness, who or what a comment is directed at, author intent, presence of specific terms
Toxic	"type of conversation widely found online which are insulting and violent in nature" (Gilda et al., 2022)	Speaker intention, presence of specific terms
Terroristic	GIFCT's Shared Industry Hash Database of terms (Gorwa et al., 2020)	Presence of specific terms
Obscene	"...the use of rude words or offensive expressions" (Rojas-Galeano, 2017, p. 12:1)	Presence of specific terms
Abusive	Speech that demonstrates disrespect toward another person (Papegnies et al., 2017), or speech designed to "personally attack others or discriminate them based on race, religion, or sexual orientation. It can also include more community-specific aspects…" (Papegnies et al., 2019)	Directedness, who or what a comment is directed at, author intent, type of behavior exhibited by author
Uncivil	"...features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics" (Coe et al., 2014)	Directedness, who or what a comment is directed at, tone

B CODEBOOK

Table 6. Codebook including descriptions of explanatory features

Feature	Description	Codes	Code Description
Directedness	Whether the text is aimed at a particular entity.	*Directed*	Comments that are aimed at a particular entity. Can be further classified into sub-categories according to the entity at which the comment is aimed: *groups*, *individuals*, or *behaviors*.
		Undirected	Comments that are not aimed at any particular entity.
Pronoun Use	The types of pronouns that text authors use in their comments.	*1st person*	Comments that use 1st person pronouns (i.e., I).
		2nd person	Comments that use 2nd person pronouns (i.e., you).
		3rd person	Comments that use 3rd person pronouns (i.e., he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, their, theirs, and themselves).
Behavior	A characterization of the actions that the text authors express via their comments.	*Name-calling*	Using a specific term to deride a particular entity.
		Insulting	Implicitly deriding a particular entity without using a specific term.
		Accusation	Insinuating that an entity acted wrongfully.
		Threat	Implicitly or explicitly suggesting that physical harm toward a particular entity would be desirable.
		Misinformation	Suggesting that commonly held beliefs are incorrect, when that suggestion could potentially lead to physically harming people.
		Benign	Behaviors that do not appear to have discernible harmful characteristics.
Communication Style	Different tones with which a commenter might express themselves. Our codes and code definitions are derived from *The Four Basic Styles of Communication*. [55]	*Passive*	When a commenter avoids expressing their opinions or feelings, protecting their rights, and identifying and meeting their needs.
		Aggressive	When a commenter expresses their feelings and opinions and advocates for their needs in a way that violates the rights of others.
		Passive-Aggressive	When a commenter appears passive on the surface but is really acting out anger in a subtle, indirect, or behind-the-scenes way.
		Assertive	When a commenter clearly states their opinions and feelings, and firmly advocates for their rights and needs without violating the rights of others.
Sarcasm	The presence of "the use of remarks that clearly mean the opposite of what they say, made in order to hurt someone's feelings or to criticize something in a humorous way." [1]	*Yes*	Sarcasm appears in the comment.
		No	Sarcasm does not appear in the comment.

C LIKELIHOOD RATIO TESTS

Table 7. Likelihood ratio test outputs for each model.

Model	#Df	LogLik	Df	Chisq	Pr(>Chisq)
H (All)	17	-318.06			
Null	2	-380.18	-15	124.24	0.0000
O (All)	17	-441.66			
Null	2	-558.48	-15	233.65	0.0000
T (All)	17	-447.15			
Null	2	-553.71	-15	213.12	0.0000
H (Only)	3	353.48			
Null	2	352.41	-1	2.13	0.1443
O (Only)	6	-47.85			
Null	2	-52.70	-4	9.71	0.0456
T (Only)	17	92.73			
Null	2	10.59	-15	164.29	0.0000
HO	3	548.23			
Null	2	545.67	-1	5.12	0.0237
HT	6	493.51			
Null	2	487.80	-4	11.41	0.0223
OT	18	-250.95			
Null	2	-301.94	-16	101.97	0.0000
HOT	19	-191.23			
Null	2	-243.56	-17	104.66	0.0000
Overlap (Any)	19	-400.52			
Null	2	-522.78	-17	244.53	0.0000

D SAMPLE CODED COMMENT

Table 8. Sample comment with codes assigned from codebook.

Comment: "@ksatnews Good for him. Glad he caught it. I do wish him well though. Maybe he will think again about sending kids to school without wearing a mask."

Feature	Code
Directedness	Directed
Directed at	Individual
Pronoun Use	3rd person
Behavior	Threat
Multiple Behaviors	No
Communication Style	Aggressive
Sarcasm	No

E SAMPLE HOT COMMENTS

Table 9. Sample comments annotated as any one of, or any combination of H, O, or T.

Concept	Sample Comment
H	@Reuters This is a genius plan to convince anti-vaxxers to get bitten by vipers.
O	@TheLeadCNN @BillKristol How the hell are you CNN people able to stomach this horse crap.
T	@globalnews @GlobalCalgary Ok Canada here's out time to kick that lying disrespected cheating no go leader out of power! So pleeeeeeeease do the right thing and don't vote liberal cause all he will do and it will happen is ruin our country and make all other countries laugh and treat us like him garbage!
HO	@JamesGRickards all of our major cities that are run by the Dims are cesspools and very dangerous. I'll stick to the road less traveled and the countryside, real America. https://t.co/wGtep8FIVN
HT	@AliVelshi They blame the people coming crossing the Border, and that didn't work, so they start on the blacks. The Republican blame game.
OT	@Jedi_MAGA @sillyfools09 @thehill No one is killing unwanted babies. It's a blob of cells, a parasite.
HOT	ALL cops that where at the capital need a kill switch....tell their truth before they get offed.

Received January 2024; revised July 2024; accepted October 2024