Operationalizing content moderation “accuracy” in the Digital Services Act

JOHNNY TIAN-ZHENG WEI, University of Southern California, USA
FREDERIKE ZUFALL, Max Planck Institute for Research on Collective Goods, Germany
ROBIN JIA, University of Southern California, USA

The Digital Services Act,[1] recently adopted by the EU, introduces new obligations for content moderation along with reporting requirements for automated systems. In it is a requirement for social media platforms to report the “accuracy” of their automated content moderation systems. The colloquial term is vague—if we are estimating accuracy (correct number of predictions divided by the total), to which ground truth are we measuring against? In addition, accuracy is not suitable for problems with large class imbalance. To operationalize “accuracy,” we relate problems of overmoderation and undermoderation to low precision and low recall, respectively. In particular, estimating recall is a challenging statistical problem. Naive estimation can incur extremely high annotation costs, which would be unduly burdensome and interfere disproportionately with the platform’s right to conduct business. Here we demonstrate, in a simulation study on the CivilComments toxicity detection dataset, that the appropriate use of stratified sampling and classifiers can greatly improve efficiency. Using this improved estimator, we study in-the-wild prevalence of personal attacks on different subreddits, and calculate realistic statistics related to efficiency, power, and minimum detectable effects. We conclude on relating these statistical concepts to guidelines for future legal clarification and implementation.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: content moderation, EU law, Digital Services Act, statistics, stratified sampling, prevalence

ACM Reference Format:
Johnny Tian-Zheng Wei, Frederike Zufall, and Robin Jia. 2023. Operationalizing content moderation “accuracy” in the Digital Services Act . 1, 1 (May 2023), 20 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The same design features that allow a social media platform to grow quickly, such as an instantaneous share button, also leave online spaces unchecked [28]. Concerns have been raised about social media platforms’ role in spreading online hate [20] and anti-vaccination misinformation [37], among many others. Research in machine learning and natural language processing (NLP) largely focuses on improving classification for the purposes of moderation [39]; fewer pieces of work examine the broader context in which moderation is conducted.

Our work identifies technical opportunities in content moderation from a legal perspective (§2). Content moderation happens between social media platforms and civilians, but a triangular relationship exists when we include the legal system into our consideration. We examine the ways legal systems can mediate their relationship (§2.1). Lawmakers and regulatory authorities are increasingly recognizing that governing content moderation is a balancing act between

[1] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (DSA)

Authors’ addresses: Johnny Tian-Zheng Wei, jtwei@usc.edu, University of Southern California, University Park Campus 3551 Trousdale Pkwy, Los Angeles, California, USA, 90007; Frederike Zufall, zufall@coll.mpg.de, Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, Bonn, Germany, 53113; Robin Jia, robinjia@usc.edu, University of Southern California, University Park Campus 3551 Trousdale Pkwy, Los Angeles, California, USA, 90007.

2023. Manuscript submitted to ACM

conflicting rights: on one hand there is a need to protect free speech rights, and on the other to curtail societal harms, such as hate speech which discriminates against minorities [13]. This balance must be made all while respecting a platform's right to conduct business.

In this context, the recently adopted EU Digital Services Act (§2.2) promises to have a far-reaching influence beyond the EU and set the standard for other legislation worldwide [8]. The Act includes a requirement for social media platforms to report the "accuracy" of their automated content moderation systems. The colloquial "accuracy" is, however, underspecified, and this is the departure point of our work: we broadly interpret "accuracy" regarding issues of free speech and societal harms (§2.3). To operationalize "accuracy," we first identify sources of underspecification, such as what defines a positive example and what test data is used (§2.4). We then relate problems of overmoderation and undermoderation to low precision and low recall, respectively (§2.5).

With a technical operationalization of content moderation "accuracy" in terms of precision and recall, the second half of our work is dedicated to recall (starting in §3). While auditing precision presents qualitative challenges, it is straightforward to estimate statistically. In contrast, auditing recall poses a challenging statistical problem, as it requires estimating the amount of undesired content over the entire platform (§3.1). Naive estimation can incur extremely high costs and, thus, reporting recall may disproportionately interfere with the right to conduct business. We show that such estimation can be made efficient by introducing stratified sampling, a method that leverages a trained classifier to efficiently estimate recall in a statistically sound and legally compliant manner (§3.2). Using the analytical formulas for stratified sampling cost, we can compare different stratification strategies (§3.4).

We first identify good strategies for stratified sampling by testing in a simulated setting using CivilComments, a toxicity detection dataset (§4). For each classifier we test with, stratified sampling outperforms random sampling, reducing the amount of annotation required by up to 64% (§4.2). The efficiency benefits are also present in measuring personal attacks on Reddit, and we provide realistic estimates of the annotation effort required (§5). With our experiments, we relate statistical concepts back to guidelines for future legal clarification and implementation (§5.2).

2 LEGAL BACKGROUND

Content moderation by social media providers may refer to deleting posts, banning or sanctioning users, or notifying authorities of potentially criminal offenses. In this section, we examine the legal context in which moderation is conducted. As legal instruments in this area are either new, voluntary, or non-existent, **our focus is in transparency**; we take the position that transparency is a necessary first step towards proper regulation.

2.1 Regulatory regimes for content moderation

Here we discuss existing and hypothetical regulatory regimes for content moderation:

Regime 1: no regulation. In the U.S., §230 of the of the Communications Decency Act[2] gives social media platforms legal protection: they are not held accountable for the content they host (with few exceptions), but are still permitted to make "Good Samaritan" edits or removal of their content. This is a massive legal privilege in hindsight, but allowed social media platforms to flourish in the early days of the internet by protecting them from lawsuits [2]. While changing §230 is technically possible, alternative proposals to regulating content moderation are extremely difficult to adopt given the strong interpretation of the Free Speech Clause of the First Amendment to the U.S. Constitution [9, 14]. Other countries

[2] 47 U.S. Code § 230.

could face similar hurdles. In France, removal obligations for content moderation[3] were deemed unconstitutional by the French Constitutional Council, as it violated the right to the freedom of expression.[4]

Even with no regulation, social media companies have strong financial incentives to voluntarily moderate their content [15]. For instance, they may wish to prevent toxic or discriminatory content from dominating their platform to attract more users and advertisers. To this end, platforms often adopt and enforce a set of “community guidelines.” Public pressure may also lead social media companies to release transparency reports [1, 21].

Regime 2: direct regulation. Given the potential harms of social media platforms, legislators have increasingly sought to enact legal requirements for certain content moderation practices. Adopted in 2017, the German Network Enforcement Law[5] obliges platforms to delete obviously illegal content when notified. Based on the example set in Germany, the EU introduced and adopted the Digital Services Act (DSA), which will be applicable from February 17th, 2024 onwards; it will apply to both providers in the EU and outside the EU, if their intermediary services are offered to recipients located in the EU (Art. 2(1) DSA). Many U.S.-based social media providers such as Twitter, Instagram or Facebook would fall under its jurisdiction. The Act introduces concrete obligations for content moderation in two categories: 1) obligations to act against illegal content and 2) transparency requirements.

Regime 3: indirect regulation. Content moderation could hypothetically be regulated indirectly through other areas of law. Competition law could be extended to establish inadequately moderating content as an unfair practice, and requiring reporting obligations from there [27]. For the U.S., this is a regulatory approach that might pass First Amendment scrutiny. Softer than direct mandatory regulation are also forms of self-regulation that may be encouraged by law, similar to Art. 40 GDPR. Self-regulating principles for content moderation, such as the Santa Clara Principles [6] and the EU Code of conduct on countering illegal hate speech online[7], have been voluntarily adopted by platforms before.

2.2 The Digital Services Act

Due to its broad impending consequences for content moderation, we now turn our focus to the Digital Services Act.

Why is the Act important? The Digital Services Act is currently the world’s most advanced legislation in content moderation. Our study of the Act is timely and crucial: governing bodies outside of the EU will likely reference the Act in considering legislation for their own jurisdictions. This is especially true in light of the regulatory power entailed by the worlds largest single market [“the Brussels effect”; 8]. Similarly to the General Data Protection Regulation (GDPR), the DSA has a broad territorial scope: it directly applies to providers outside the EU (Art. 2(1) DSA).

Which social media platforms will be affected by the Act? The Act introduces a nested classification of online platforms, with additional specificity and obligations for each nested level. At the highest level are the “intermediary services,” an umbrella term subsuming “hosting services,” which subsumes “online platforms,” which subsumes “very large online platforms” (VLOPs). All obligations for “intermediary services” apply to “online platforms,” as online platforms provide hosting services at the request of the recipient and store and disseminate information to the public (Art. 3 (i) DSA). Intermediary services that provide to more than 45 million average monthly active recipients in the Union (Art. 33(1)

[3] Loi n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet.
[4] Décision n° 2020-801 DC du 18 juin 2020.
[5] Netzwerkdurchsetzungsgesetz vom 1. September 2017 (BGBl. I S. 3352).
[6] https://santaclaraprinciples.org/
[7] https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985

DSA), which likely includes U.S.-established platforms Twitter, Instagram, and Facebook, are classified as "very large online platforms" (VLOPs) and will have to comply with the full set of requirements set forth in the Act.

What moderation requirements are in the Act? While there is no general obligation to monitor information (Art. 8 DSA), the Act obliges intermediary services to act against illegal content upon order by a judicial or administrative authority (Art. 9 DSA), or upon notice by an individual or entity (Art. 16, Art. 6(1)(b) DSA). The Act also acknowledges the right to voluntarily detect, remove, or disable access to illegal content (Art. 7 DSA). Providers may use content moderation, automated or manual, to detect illegal content or information incompatible with their terms and conditions (Art. 3 (t) DSA). The notion of "illegal content" is rather broad, comprising content "not in compliance with Union law or the law of any Member State [. . .] irrespective of the precise subject matter or nature of that law" (Art. 3 h) DSA)).

2.3 Interpreting "accuracy" as a transparency requirement

What transparency requirements are in the Act? The DSA stipulates that Terms and Conditions must contain information on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic decision-making and human review (Art. 14(1) DSA). Even further, providers are obliged to publish comprehensive reports at least once a year on any content moderation they engaged in during the relevant period (Art. 15(1) DSA). VLOPs have an even increased reporting obligation: they have to publish these reports every six months, and include, amongst others, information on human resources engaged with content moderation (Art. 42(1),(2) DSA). These reports must include not only the number of orders received from authorities and notices submitted by users to the platform, but also information on *voluntary* content moderation by the provider (Art. 15(1)(a),(b),(c) DSA). In all these cases, Art. 15(1)(e) DSA requires providers to report:

> "any use made of automated means for the purpose of content moderation, including a qualitative description, a specification of the precise purposes, **indicators of the accuracy and the possible rate of error of the automated means** used in fulfilling those purposes, and any safeguards applied."

In addition, VLOPs must also conduct mandatory risk assessments which include their content moderation systems. These assessments must be communicated to the EU Commission or the Digital Services Coordinator upon request (Art. 34(2)(b) DSA). The assessment is required to focus on balancing the risks between curtailing the spread of illegal content on their platform with a potential infringement on fundamental human rights. These rights may include the right to the protection of personal data (Art. 8 EU Charter[8]) and freedom of expression, as enshrined in Art. 11 EU Charter (Art. 34(1)(a),(b), Art. 35 DSA).

How can we understand what is meant by "accuracy"? Troublingly, the DSA leaves the term "accuracy" underspecified. Art. 15(1)(e) DSA only asks for *"indicators of the accuracy and the possible rate of error"* (Art. 15(1)(e) DSA). From a technical perspective, the term "accuracy" is insufficiently specific. If we are estimating accuracy (correct number of predictions divided by the total), to which ground truth are we measuring against? In addition, accuracy is not suitable for problems with large class imbalance. As it stands, "accuracy" reporting required by the DSA will exhibit inconsistency. Social media providers may each make their own decisions regarding test data and evaluation metrics.

Here we try to elucidate the notion of "accuracy" in Art. 15(1)(e) of the DSA using a method of interpretation in EU law: inter-instrumental interpretation [18]. By referencing the definitions and systematic understanding from another instrument of EU law, we can make an attempt to apply it to "accuracy." We refer to the recent proposal for an Artificial

[8] Charter of Fundamental Rights of the European Union, OJ C 326, 26.10.2012, p. 391.

Intelligence Act (AI Act),[9] a parallel development in EU law related to specific high-risk automated systems. Similar to the DSA, the AI Act would apply to U.S.-based providers if they are "placing on the market or putting into service AI systems in the Union" (Art. 2(1)(a) AI Act). While the AI Act is specifically targeted towards high-risk systems, it outlines a regulatory approach to classification systems. Thus, the AI Act is suitable for inter-instrumental interpretation.

With respect to "accuracy", Art. 15(1) of the AI Act proposal requires "(high-risk) AI systems to be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of *accuracy* [...] and perform consistently in those respects throughout their lifecycle." Notably, the draft only mentions "accuracy" and not other metrics such as precision or recall. Upon further inspection, however, the legislatory process reveals that the draft was inspired by the policy guidelines of the High Level Expert Group on AI [33]. In the guidelines, the use of accuracy is intended to reflect the "trustworthiness" of an AI system, but a complementary footnote stated that "accuracy is only one performance metric and it might not be the most appropriate depending on the application," and that the F1 score, false positives, and false negatives may also be used.

Our search concludes in identifying an oversight in EU law, where "accuracy" is underspecified. At the same time, this presents an opportunity—the following discussion operationalizes "accuracy" in the context of content moderation.

2.4 What are the true positives? On which test data?

If the intent of the DSA is for companies to report the "accuracy" (i.e. the performance, trustworthiness, or reliability) of content moderation, any technical evaluation will need three basic specifications. First, the *ground truth*, or what content should legitimately be removed, will need to be determined. The DSA addresses two classes of content in its definition of "content moderation" (Art. 3(t) DSA):

- *Illegal content.* Illegal content is defined as "any information that [...] is not in compliance with Union law or the law of any Member State [...], irrespective of the precise subject matter or nature of that law" by Art. 3(h) DSA. For instance, national law of the EU Member States must contain a minimum standard to incriminate hate speech as specified by an EU Framework Decision.[10] Furthermore, the EU Commission has also started an initiative to add "all forms of hate crime and hate speech, whether because of race, religion, gender or sexuality"[11] to the list of EU crimes in Art. 83(1) TFEU[12]. To implement the detection of illegal content (e.g., punishable hate speech) in practice, however, would require an operationalization of the definition to identify the class of illegal content [46]. To be feasible, such identification must be possible for a skilled person with no legal training because of the labor costs.
- *Content that violates community guidelines.* Another content moderation scenario that the DSA specifically addresses are any activities that are aimed "at detecting, identifying and addressing [...] information incompatible with their terms and conditions" (Art. 3(t) DSA). An advantage in identifying whether speech violates community guidelines is that it is already operationalized by the platform. The measurement of accuracy on the guidelines is a measurement of whether the moderation system is achieving its intended purpose. This would also complement other transparency measures about the guidelines themselves (see Art. 14 DSA). However, the guidelines offered by online platforms may not be founded in any democratic or legal basis, and may not even be politically neutral

[9] COM(2021) 206 final.

[10] Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law and national laws transposing it.

[11] Communication of 9.12.2021, COM(2021) 777 final. This would allow the Commission to replace the existing Framework Decision by a new Directive further elaborating on a more extensive notion of hate speech incrimination.

[12] Treaty on the Functioning of the European Union, [2016] OJ C202/1.

[44]. Guidelines may still be subject to judicial redress if they interfere with protected rights of the respective users, such as (but not limited to) the freedom of speech.[13]

Second, the *test data*, or dataset on which we seek to evaluate the classification performance, needs to be considered. Since the DSA specifies reporting periods of six months or one year, it would be natural to evaluate on content accumulated within the last reporting period. We speculate on two sets of data appropriate for reporting:

- *All visible content.* An accuracy metric could be evaluated over all visible content on the platform.
- *All notified content.* Art. 16 DSA requires providers to put in place mechanisms to allow users to notify (i.e., flag) content they consider to be illegal. If these notices allow identifying illegality without a detailed legal examination, they create a liability for providers if they do not act expeditiously to remove or disable access to this content (Art. 16(3), Art. 6(1)(b) DSA). This notified content may be a subset of interest. Reserving analysis to the notified set reduces the total amount of data that must be considered for evaluation. However, anyone can flag content without concrete reason, possibly even with malicious intent [29].

For the rest of this paper, we will consider the evaluation of content moderation with respect to **community guidelines** on **all visible content**. These settings are chosen so that academic study is possible, but our methods are general.

2.5 Choosing metrics to disclose

The final consideration is which *evaluation metric* to disclose, which we now discuss in detail.

Accuracy. This metric is by far mentioned the most in EU law. Its use in the Digital Services Act appears to be colloquial and refers to classification reliability in general. However, we provide the technical definition here:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, FP, TN, FN are true positives, false positives, true negatives, false negatives, respectively. Accuracy can be highly inflated if the test data is label-imbalanced, as a classifier that only outputs the majority label will achieve high accuracy. Content moderation is a label-imbalanced problem, as legal content greatly outnumbers illegal content. Under a strict interpretation of the Act, the use of accuracy would be technically inappropriate. Instead, we provide our interpretation of "accuracy" as an indicator of content moderation reliability through precision and recall.

Precision. This metric represents one side of moderation "accuracy" and measures overmoderation:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2}$$

Precision measures the fraction of examples positively predicted (e.g., removed by a content moderation system) that are truly positive (e.g., content that is illegal or violates community guidelines, depending on the ground truth). Therefore, precision is appropriate for tracking overmoderation: low precision indicates that content that is not considered removable, which may include free speech, is being removed.

Recall. This complementary metric represents another side of moderation "accuracy" and measures undermoderation:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3}$$

[13] See the case law from the German Constitutional Court, 22.05.2019, ECLI:DE:BVerfG:2019:qk20190522.1bvq004219.

and measures the fraction of truly positive content (e.g., illegal posts) that was predicted positive (e.g. removed by the moderation system). Recall is appropriate for tracking undermoderation: low recall indicates that that a lot of removable content, which may include illegal content, is still visible on the platform.

Trade-offs and F1-score. The closest metric in EU law to the ones mentioned here is the F1 score, which is the harmonic mean between precision and recall. However, we encourage legislators and regulators to reason about content moderation "accuracy" in terms of precision and recall separately, as these relate to problems of overmoderation and undermoderation, respectively. Note that there is a tradeoff between precision and recall: most automated classification systems have a decision threshold that can be tuned to overpredict or underpredict positives. If content moderation systems are tuned for overprediction (i.e. remove more content), it can be reasonably expected that recall (and overmoderation) increases and precision (and undermoderation) decreases, and vice versa for underprediction.

Auditing and improving precision and recall. Auditing content moderation precision is a qualitative undertaking. To audit precision, removal decisions will need to be examined closely, preferably by a lawyer. Improving precision may involve training moderators to make better judgments [38]. Providing users the right to judicial redress (i.e., allowing users to contest removal decisions) can also improve precision by revising decisions on borderline speech.

High recall is achieved when the visible content on the platform is mostly "clean." The straightforward approach to improving recall is to apply more labor hours for direct moderation. Platforms may also try interventions to discourage users from posting bad content, such as disbanding toxic communities [41]; recall could be used to evaluate such interventions. However, measuring recall poses a challenging statistical problem, to which we now turn our attention.

3 ESTIMATING RECALL

We now focus on the estimation of recall and its related quantities. Exactly computing recall for content moderation is highly nontrivial, as a large platform's content cannot feasibly be exhaustively annotated. This same challenge is well known for tasks like information retrieval, as recall is defined over all webpages on the web [3]. Similarly, recall for knowledge base population is difficult to compute since it is defined over all relational facts in a document collection [10]. The only practical alternative is to *statistically estimate* recall via sampling and targeted annotation.

The reporting obligation required by the Digital Services Act must take into consideration the annotation effort needed from the platforms to fulfill this reporting requirement. A burdensome recall reporting requirement may interfere disproportionately with the platform's freedom to conduct a business, as enshrined in Art. 16 of the EU Charter. Here we explore the use of stratified sampling, a variance reduction technique that can reduce the cost of estimation, by combining it with NLP classifiers.

Crucially, stratified sampling yields statistically unbiased and consistent estimates: the estimate's expected value equals the true recall, and the estimate converges to the true recall given enough annotated samples [34]. For rigorous evaluation, recall estimation must be *unbiased* against human judgment. In contrast, fully automated prevalence estimation [23] to estimate moderation recall is a *biased* strategy and would not be rigorous, as automated classifiers are known to make many errors relative to human annotated ground truth [17, 45]. We reason that legal justification will also require unbiased estimates. Stratified sampling is unbiased even if the leveraged classifier has undesirable properties such as low accuracy or low fairness [7], as a human is the final decision maker. A suboptimal classifier may cause the method to require more annotations, but it cannot introduce bias.

3.1 Setting up the estimation problem

We consider the goal of estimating recall to measure whether a content moderation system is achieving its intended purpose, i.e., capturing all content that violates community guidelines. We assume that we know the amount of violating content removed (true positives), as this can be estimated directly by inspecting the set of removed content. Estimating recall thus requires estimating the amount of violating content that remains visible on the platform (false negatives). We focus on estimating the probability of violating content, or prevalence:

$$p = \mathbb{E}_{x \sim \mathcal{U}}[s(x)] \tag{4}$$

where $s(x) = \mathbb{I}(x \in S)$ is an binary function indicating whether the content x violates community guidelines (i.e., ground truth), and $\mathcal{U}$ is the pool of content not removed and visible on the platform (i.e., predicted negatives) during the last reporting period. If we had the probability p, we could use $FN = p|\mathcal{U}|$ to obtain the number of false negatives.

The simplest way to estimate this probability is by applying the random sampling estimator $\hat{p}$ where

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} s(x_i) \tag{5}$$

and $\{x_i, i = 1...n\}$ is a random sample from the remaining content $\mathcal{U}$. This estimator has standard error $\mathrm{SE}_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$, which is directly proportional to width of the confidence interval on p if a normal approximation is applied. This quantifies how precisely we have estimated p. Typically, the prevalence is rare e.g. $0 < p << 0.1$. When estimating the probability of a rare event, we would like to control the coefficient of variance [12],

$$\mathrm{CV}_{\hat{p}} = \frac{\mathrm{SE}_{\hat{p}}}{\hat{p}}, \tag{6}$$

which is the relative estimation precision to the prevalence. In other words, we are estimating the rare probability within some percentage accuracy. If we require estimates to have a $\mathrm{CV}_{\hat{p}} \leq 20\%$, the rarer the true prevalence, the smaller the standard error will need to be.

3.2 Stratified sampling

Given any partition of the data into strata (i.e., bins), we can apply the stratified sampling estimator. This estimator is computed by annotating a random sample within each stratum, then combining the mean for each stratum in an unbiased manner. A good stratification can significantly improve the variance of the stratified sampling estimator over random sampling. In particular, if some strata have very low or very high prevalance, those strata will have lower variance. Then, we can allocate our samples to focus on estimating prevalences of the other high variance strata.

We adopt notation from [32]. Let L be the number of strata, N be the total size of the dataset, N_h be the number of total items in stratum h, and n_h be the number of samples drawn and annotated from stratum h for estimation. The stratified sampling estimator $\hat{p}_{st}$ is defined by the equations:

$$\hat{p}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \cdot \hat{p}_h \qquad \text{where} \qquad \hat{p}_h = \frac{1}{n_h} \sum_{n_h}^{i=1} s(x_i^h) \,\forall\, h \in \{1, \ldots, L\}; \tag{7}$$

and $\{x_i^h\}_{i=1}^{n_h}$ is a random sample of n_h examples from stratum h. In other words, $\hat{p}_{st}$ is a weighted average of the $\hat{p}_h$'s, which are our prevalence estimates within each stratum. This estimator is unbiased, i.e., the expected value $\mathbb{E}[\hat{p}_{st}]$ of the estimator equals the true probability p.

Given a stratification, we can analytically calculate the variance of the stratified estimator as

$$\widehat{\mathrm{Var}}(\hat{p}_{st}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right)^2 \cdot \frac{\hat{p}_h(1-\hat{p}_h)}{n_h}. \tag{8}$$

Note that the second term $(1 - n_h/N_h)$ is a finite population correction and ≈ 1 when the strata are large relative to the number of annotations. The standard error is then $\mathrm{SE}_{\hat{p}_{st}} = \sqrt{\widehat{\mathrm{Var}}(\hat{p}_{st})}$ and the confidence intervals of the estimator can be calculated by assuming the estimator follows a normal distribution (by the central limit theorem).

3.3 Binning and allocation

The variance of stratified sampling depends on two choices: binning (i.e., how to partition the examples) and allocation (i.e., how many examples to annotate in each stratum). We now describe methods to choose each.

Binning. An effective way to create strata is with a classifier trained to identify violating content. Given a piece of content, the classifier predicts the probability that it violates community guidelines. With the predictions scores from the classifier, content can be binned in a few ways by their score. The number of bins L is a hyperparameter, which we investigate in §4. **(1)** *Equal width binning.* The classifiers we consider score each comment from $[0, 1]$. Equal-width binning splits this interval into L equal width intervals each corresponding to a bin. Sampling across bins thus samples across the entire scoring range of the classifier. **(2)** *Quantile binning.* Once the comments are scored, quantile binning splits the scoring range into equal sized quantiles, where each bin will contain the same number of points. The range of scores within each quantile will depend on the scoring distribution of the classifier. **(3)** *Oracle binning.* This binning is not possible in practice as it requires the labels of all the data in the pool, but we include it as an oracle method. Using a recursive search procedure and the labels of all the training data, we can search for a good binning. At each recursive step, a bin is split into two at a point where the resulting variance is minimized. We recursively break down the classifier’s scoring interval into a number of bins that is a power of two.

Allocation. Once the bins are partitioned, an allocation of how to sample from each strata (i.e., a choice of the n_h’s) can be made. **(1)** *Equal allocation.* This baseline allocation samples from each bin equally. **(2)** *Optimal allocation.* For a given stratification, the optimal sample allocation minimizes the variance of the estimator. For each stratum, the number of samples to allocate is given by

$$n_h^{opt} = n \cdot \frac{N_h \sigma_h}{\sum_k N_k \sigma_k} \tag{9}$$

where σ_h is the standard deviation of samples within stratum h, and n is the total number of planned samples to annotate. This allocation is an oracle method: it cannot be run in practice, as it requires the standard deviations within each stratum, which are not known beforehand. We include this allocation method for analysis purposes. **(3)** *Pilot allocation.* To approximate the optimal allocation, we adopt a simplified version of Bennett and Carvalho [5]. We annotate a fixed number of pilot samples within each stratum, which we use to estimate the standard deviation per stratum. We then approximate the optimal allocation using these standard deviations. If the size of the pilot sample within the stratum exceeds the optimal allocation for that stratum, no additional samples are collected for that stratum. If the optimal allocation is more than the pilot sample, the additional samples are collected. We use psuedocounts (adding 1 positive and 1 negative example) to ensure each stratum is allocated samples.

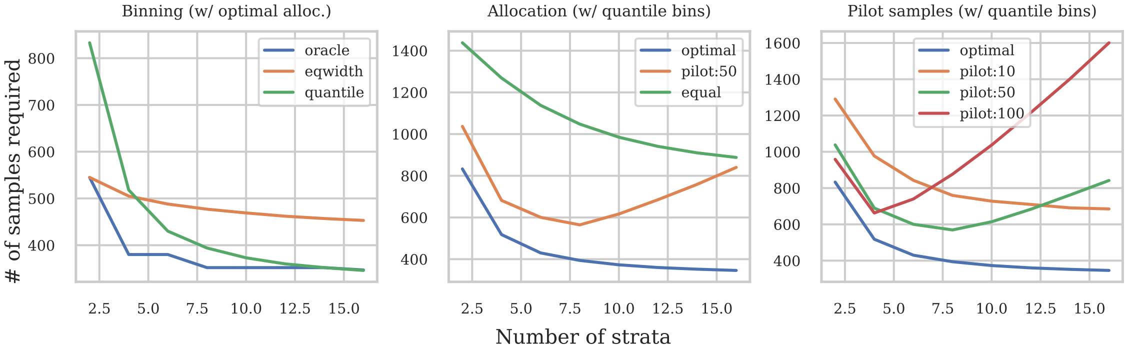

Fig. 1. Number of samples required to estimate the prevalence of toxicity (5.9%) within 20% in CivilComments, with different number of strata. Binning is based on predicted scores from a finetuned Roberta classifier. All pilot results are averaged over 30 trials. As the number of strata grows, estimators using optimal allocation require fewer samples. Pilot methods outperform the equal allocation baseline, but when there are too many strata, annotation effort is wasted on the pilot samples.

3.4 Analyzing the burden of estimation

The variance of the random and stratified sampling estimators can be calculated analytically, so their power (i.e. number of samples needed to achieve some estimation accuracy) can also be calculated if we assume the central limit theorem. This assumption is justified because the sample sizes we consider are large, so the estimators are normally distributed.

We consider the goal of reporting $\hat{p} \pm (0.1)\hat{p}$ with 95% probability. The coefficient of variance is effectively required to be $\text{CV}_{req} = 0.1/z_{.95}$, and in turn $\text{SE}_{req} = (0.1)\hat{p}/z_{.95}$, where $z_{(1-\alpha)}$ is test statistic of the normal variable. Since SE_{req} is a function of n, the total number of annotations collected, we can solve for n for each estimator. For random sampling we have

$$\text{SE}^2_{req} = \frac{\hat{p}(1-\hat{p})}{n} \qquad \therefore n = \frac{\hat{p}(1-\hat{p})}{\text{SE}^2_{req}} \tag{10}$$

to achieve the desired accuracy. A table of power calculations for random sampling is given in Appendix A.

Given a stratification, we can analytically calculate costs as well. For equal allocation stratified sampling, we have

$$\text{SE}^2_{req} = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \cdot \frac{\hat{p}_h(1-\hat{p}_h)}{(1/L)n} \qquad \therefore n = \left(\frac{L}{\text{SE}^2_{req}}\right) \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \cdot \hat{p}_h(1-\hat{p}_h) \tag{11}$$

where we simplified the variance term of the stratified estimator by dropping the population correction (if N_h is large relative to n, the correction is approximately 1). For optimal allocation stratified sampling, we have

$$\text{SE}^2_{req} = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \cdot \frac{\hat{p}_h(1-\hat{p}_h)}{n_h^{opt}} \qquad \therefore n = \left(\frac{1}{\text{SE}^2_{req}}\right) \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \cdot \frac{\hat{p}_h(1-\hat{p}_h)}{c_h^{opt}} \tag{12}$$

where c_h^{opt} is the proportion of n allocated to stratum h. Using these analytical cost calculations, we can empirically calculate the expected cost for different sampling methods on different data, by filling in the appropriate observations.

	Binning	Allocation	20%	10%	5%
random sampling	n/a	n/a	1532	6127	24508
unigram	oracle:8	optimal	559	2235	8937
	quantiles:8	pilot:50	716 ± 23	2868 ± 94	11461 ± 366
roberta (balanced)	oracle:8	optimal	353	1410	5639
	quantile:8	pilot:50	567 ± 15	2207 ± 82	8815 ± 304
distilbert (balanced)	oracle:8	optimal	378	1510	6038
	quantile:8	pilot:50	574 ± 18	2246 ± 81	9009 ± 346

Table 1. Number of samples needed for the stratified sampling estimator to estimate prevalence (5.9%) in civilcomments within a certain percentage. Oracle:8 means 8 strata were created and pilot:50 means 50 pilot samples are taken within each strata for variance estimation. Oracle binning and optimal allocation are oracle techniques which rely on labelled data that is not available beforehand in practice. We chose 8 bins and 50 pilot samples because these settings were most effective in Figure 1.

4 SIMULATING STRATIFIED SAMPLING

We now conduct an empirical study of stratified sampling for estimating prevalence (from which we can estimate recall) for content moderation. We determine which choices for binning, allocation (§3.3), and classifier are most effective by simulating it on a labelled dataset. Similar to pool-based active learning [40], we assume that the dataset has no labels, and “annotate” an example by revealing the label from the dataset. We use civilcomments [7] as our testbed for toxicity prevalence. While toxicity is not an operationalized legal definition or community guideline, this setting provides a large dataset to study sampling in general.

4.1 Experimental setup

Dataset. The civilcomments dataset [7] contains comments labeled by their toxicity, from the archive of the Civil Comments platform, which is a commenting plugin for independent news sites. Each comment is annotated for toxicity by at least 10 annotators, and if half of the annotators label it toxic, the toxicity label is positive. For our study, we use the test set for training and the training set for simulation, because the training set (1.8M examples) is much larger than the test set (97K examples), which better represents content moderation settings where $\mathcal{U}$ may be large and the number of labeled examples is relatively small. This dataset naturally presents a rare class prevalence estimation problem as toxic comments have a prevalence of 5.9%.

Models. We experiment with a few supervised classification techniques. These classifiers are used to score the unlabeled examples, from which the strata are created. We study three models: a unigram logistic regression model, Roberta [25] (a pretrained transformer model), and Distilbert (a pretrained transformer model, 40% smaller than Roberta). Training details and hyperparameters are in Appendix B.

4.2 Results

As shown in Figure 1, using the right techniques and parameters for stratified sampling makes a large difference: the worst choices perform as poorly as the random sampling estimator, while the best choice can save up to half of the annotation effort. With optimal allocation, increasing the number of strata decreases variance. Intuitively, optimal allocation is more effective with more bins because a more precise allocation is given to each strata, according to the strata prevalence. The pilot allocation, which approximates the optimal allocation, outperforms equal allocation. However, as the number of bins increases, the total number of pilot samples also increases, resulting in wasted samples.

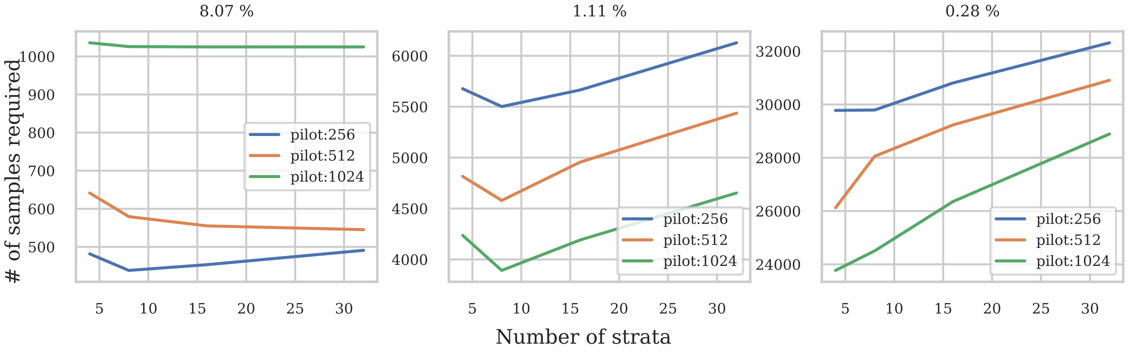

Fig. 2. Number of samples required to estimate the prevalence of toxicity within 20% in civilcomments, with different prevalences, with different number of strata. Binning is based on predicted scores from a finetuned Roberta classifier. All pilot results are averaged over 100 trials. The numbers associated with the pilot is the *total* number of pilot samples across all bins. For small prevalences, more pilot samples are better, because more samples estimate the standard deviations within each bin better. In contrast, more bins fares worse because each bin has less pilot samples.

In Table 1, we show that using a classifier can greatly reduce annotation requirements over a random sampling estimator. Estimating prevalence within 20% of 5.9% requires only a few hundred samples. While the unigram model is effective, Roberta dramatically reduces the number of samples needed to estimate the prevalence. The gap between quantile and oracle binning methods is larger for Roberta than for unigram models, suggesting that Roberta would benefit from a better allocation method.

Recommendations. Based on our simulation in civilcomments, our recommendations for conducting stratified sampling in practice are: **(1)** Choose quantile binning. Among the two practical binning methods, quantile binning outperforms equal width binning. It performs at near oracle binning performance if the number of bins is sufficiently large and the optimal allocation is used. **(2)** Choose pilot allocation. Attempting to estimate a good allocation with pilot samples will greatly reduce the variance of the stratified sampling estimator. **(3)** Make a reasonable choice for the number of pilot samples. If the prevalence is suspected to be low, or the estimation requirement is precise, inefficiency in pilot samples is less of a concern (because the pilot samples are needed anyways). When the prevalence is high and the number of samples required may be less than the total pilot samples, a reasonable choice needs to be made about the number of pilot samples collected. The experimenter must consider how much data annotation is acceptable at worst. **(4)**

Rarer prevalences. For each example, civilcomments has a real valued label of toxicity corresponding to the proportion of annotators that labelled it as toxic. When converting the labels to binary, a threshold of 0.5 yields a prevalence of 5.9%, which is what we report on in §4. Here we present results for different prevalences by varying the threshold among $\{0.48, 0.83, 0.95\}$. We vary the number of total pilot samples and bins, and by choosing appropriate powers of two, each bin will always have an integer number of pilot samples.

Refer to Figure 2. In general, stratified sampling provides significant efficiency over random sampling (comparing to the figures in Table 4). The number of pilot samples collected matters for both very small and very large prevalences. In the case of high prevalence, pilot samples outnumber the optimal allocation so there will be wasted samples. In the case of low prevalence, additional pilot samples help estimate the standard deviation of the strata correctly for optimal allocation. Finally, increasing the number of strata does not help with lower prevalences, as less samples are available for each bin to estimate the standard deviation with.

5 “IN THE WILD” PREVALENCE ESTIMATION

This study aims to demonstrate what a recall reporting requirement could look like under the Digital Services Act. We focus on the moderation recall of personal attacks on Reddit,[14] which has community guidelines against such speech. Moderation recall, which is determined by the prevalence of violations, has been used to understand the effectiveness of content moderation policies. In Karsten [22], the toxicity of German political accounts was monitored before and after NetzDG was enacted; this was done completely with automated judgments, and thus may be biased. A more rigorous study could involve operationalizing a few categories of illegal or inflammatory speech, and applying stratified sampling with human annotation to obtain unbiased estimates. Srinivasan et al. [43] study whether users make fewer comments that violate community guidelines after having their comments removed.

Social media research uses prevalence in online discussions to study the causal effect of real world events, assuming that prevalence online is a proxy for real world phenomena such as public opinion [31]. Siegel et al. [42] attempts to understand how the 2016 U.S. presidential election affected hate speech online. They compute biased prevalences based only on simple classifiers, at the same time noting that simple classifiers often make mistakes. The rigor of their results could be improved by using the stratified sampling estimator we propose here. Park et al. [35] provides estimates for the prevalence of macro-norm violating comments (violations of implicit Reddit site-wide rules derived from Chandrasekharan et al. [11]). They use stratified sampling with two strata, using a classifier for stratification. Our results provide insights in applying and improving their estimation.

5.1 Experimental setup

While publicly viewable posts from Reddit are accessible through the Pushshift data stores, records for comments removed by the moderators are incomplete. The number of true positives could be easily estimated by Reddit (as they would have access to the removed comments), so we turn our focus to the prevalence of false negatives. By collecting pilot annotations for a number of subreddits, we can get a preliminary estimate on the prevalence, the number of samples needed for accurate estimation, and the number of samples needed to set up estimators with some minimum detectable effect. To conclude, we relate these statistical concepts to guidelines for future legal clarification and implementation.

Study of interest. We restrict our study to a recent dump of Reddit (December 2022) on Pushshift.io [4]. Reddit is an online social media platform organized into smaller communities (subreddits) where people share and comment on posts. Most of the discussion on Reddit happens within the comments on each post, so we evaluate the personal attacks in the comment section. We select 6 subreddits with a range of macro-norm violation prevalence, as estimated in Park et al. [35], and that have more than 1000 comments per day[15]:

(most macro-norm violations) r/politics, r/AskReddit, r/sex, r/pcmasterrace, r/wow, r/legaladvice (least)

We evaluate the prevalence of personal attacks in the non top-level comments. At the time of writing, 4 out of the 6 subreddits we study explicitly have rules against "personal attacks" (politics: rule 4, AskReddit: rule 8, wow: rule 1, legaladvice: rule 5). The other two subreddits: r/sex, r/pcmasterrace, also have rules regarding constructive engagement and respectful conversation, respectively.

[14] https://reddit.com/

[15] Accessed January 21st, 2023, as per https://subredditstats.com/.

			n needed for $\text{CV}_{\hat{p}_{st}} \leq 20\%$		Strat. samp.	1% MDE
	$\hat{p}_{st}$	95% CI	Stratified	Random	efficiency	n (= var.)
r/politics	7.00%	[6.42%, 7.58%]	[740, 1033]	[1171, 1401]	26%	6801
r/AskReddit	4.75%	[4.29%, 5.21%]	[968, 1425]	[1748, 2142]	33%	4198
r/sex	3.00%	[2.63%, 3.37%]	[1508, 2477]	[2754, 3556]	30%	2741
r/pcmasterrace	4.00%	[3.61%, 4.39%]	[1003, 1488]	[2090, 2567]	42%	3097
r/wow	4.25%	[3.76%, 4.74%]	[1314, 2078]	[1933, 2455]	15%	4711
r/legaladvice	1.50%	[1.25%, 1.75%]	[2561, 5027]	[5391, 7591]	33%	1256

Table 2. Statistics derived from pilot annotations. Stratified sampling was applied with 8 quantile bins, 50 pilot annotations per bin, and a Roberta finetuned on personal attack data. Estimating the prevalence here is equivalent to using equal proportion allocation. The number of samples required for estimation with random and pilot stratified sampling is provided assuming either the upper or lower bound of prevalence. Efficiencies are calculated assuming the lower prevalence (the pessimistic efficiency). The number of samples required to achieve a minimum detectable effect (MDE) assumes the other estimator compared has equal variance.

			n needed for $\text{CV}_{\hat{p}_{st}} \leq 20\%$		Strat. samp.	1% MDE
	$\hat{p}_{st}$	95% CI	Stratified	Random	efficiency	n (= var.)
r/politics	7.00%	[6.42%, 7.58%]	1033	1401	26%	6801
r/AskReddit	4.75%	[4.29%, 5.21%]	1425	2142	33%	4198
r/sex	3.00%	[2.63%, 3.37%]	2477	3556	30%	2741
r/pcmasterrace	4.00%	[3.61%, 4.39%]	1488	2567	42%	3097
r/wow	4.25%	[3.76%, 4.74%]	2078	2455	15%	4711
r/legaladvice	1.50%	[1.25%, 1.75%]	5027	7591	33%	1256

Table 3. Statistics derived from pilot annotations. Stratified sampling was applied with 8 quantile bins, 50 pilot annotations per bin, and a Roberta finetuned on personal attack data. Estimating the prevalence here is equivalent to using equal proportion allocation. The number of samples required for estimation with random and pilot stratified sampling is provided assuming either the upper or lower bound of prevalence. Efficiencies are calculated assuming the lower prevalence (the pessimistic efficiency). The number of samples required to achieve a minimum detectable effect (MDE) assumes the other estimator compared has equal variance.

Data collection. To download the comments, we first randomly sample a post submission (akin to a thread) from the subreddit in the Pushshift data dump. We then use Praw[16] to collect all the non-top level comments in the submission. We repeat this process until we have either 100K comments or have exhausted all possible comments for that subreddit.

Personal attacks. In our preliminary analysis of comments (Appendix C), we found that personal attacks were easiest to identify consistently. This is corroborated by Habernal et al. [19], which reports that personal attacks in r/changemyview have high inter-annotator agreement. Using the data from Habernal et al. [19], we trained a new Roberta classifier on a balanced training set with r/changemyview personal attacks and an equal number of negatives from the dataset. We label a comment as a personal attack if it fits any of criteria defined in [19, Table 2]. See Table 5 for a sample. We found most personal attacks to be immediately recognizable because it included insults or accusations. For borderline cases, such as rude comments, we labelled it a personal attack if it didn't engage with the parent comment. In some other cases, viewing the original comment thread of Reddit was helpful. When the comment was part of a "flame war", we would annotate it as a personal attack. All annotation was conducted by the first author, and the rate of annotation was about 600 examples per hour (other annotation details are in Appendix C).

[16] https://praw.readthedocs.io/en/stable/

5.2 Results

Refer to Table 3. As a sanity check, we see that ranking subreddits by personal attacks aligns closely with their ranking by macro-norm violation. By the prevalences, we can conclusively determine that r/politics has the biggest problem with personal attacks out of all the subreddits we examine. This may be because personal attacks naturally occur more in r/politics, or due to a lack of moderator effort, which could be studied with internal Reddit data [24].

As the prevalence gets smaller, the number of samples required to estimate the prevalance within some percentage of the prevalence grows larger. This is because the variance of our estimate needs to be much smaller in absolute value. For prevalences that are smaller than one percent, estimating prevalence accurately is difficult. However, a 1% absolute change in prevalence becomes easier to detect (the minimum detectable effect, or MDE). Intuitively, 1% becomes relatively large compared to the rare prevalence. The efficiency of the stratified sampling estimator (over random sampling) differs across subreddits, with the estimator performing the worst on r/wow. In r/wow, the highest scoring bin did not contain the greatest number of personal attacks (the second highest bin contained the most), unlike other subreddits. Future work may involve using active learning to improve the classifier on the fly [26, 40].

We make the following recommendations for future legal clarifications and implementations of the accuracy reporting requirement: **(1)** Smaller prevalences are harder to measure. Arbitrarily small prevalances cannot be accurately quantified. However, smaller prevalences also mean that this class of content is rare and less likely to cause major harm. It may be appropriate to set a threshold under which is deemed safe and unnecessary to estimate accurately. **(2)** Estimate how much annotation is needed. With the normal approximation and the pilot samples, we are able to make a pilot estimation on many of important experimental parameters such as a pilot estimate of prevalence and number of samples needed to estimate within some accuracy. **(3)** Plan for a minimum detectable effect [MDE; 6]. It may be useful to decide on an MDE, say 2%. If we make an equal variance assumption on the two compared estimators, we can plan for enough examples to be collected so that we may detect the specified MDE. If the true prevalence increases from one reporting period to another by more than the MDE, a statistically significant comparison will be observed (non-overlapping confidence intervals), akin to an “alarm.” An added advantage of setting up an MDE is that its cost decreases as the prevalence gets smaller, opposed to increasing when estimating prevalence and controlling the coefficient of variance.

6 CONCLUSION

In this work, we identified an oversight in the Digital Services Act regarding the reporting of “accuracy” for content moderation. The colloquial “accuracy” does not elucidate the technical implementation and will need clarification through administrative guidelines or legislative amendment. This process should be a collaborative effort including lawmakers, administrators, and computer scientists to ensure that any reporting requirements are both legally and technically sound. Likewise, a vague notion of “accuracy” is also present in the recently proposed AI Act, where it is used to refer to the performance of high-risk AI systems. While our discussion and hypotheticals are specific to content moderation, our main point applies more broadly: “accuracy” needs clarification in the AI Act as well. We hope that future work on the AI Act will find our argumentation a useful reference.

In governing content moderation, recall and precision are more appropriate metrics than accuracy. Low precision signals an issue of overmoderation, given that an appropriate ground truth can be provided; low recall signals undermoderation, given that it is computed on an appropriate subset of content. Requiring the reporting of precision and recall will help monitor issues related to protecting free speech and curtailing societal harms, respectively. With respect to social media platforms, we hope that lawmakers take into consideration the notions that reporting requirements should be structured so that they serve as useful evaluations for the platform themselves, and that the burden of reporting should not interfere with their right to conduct business.

Reporting recall poses a challenging statistical problem. Legal considerations in respecting the rights of all entities involved posed a unique set of constraints on the technical methodology. We introduced stratified sampling, a method that simultaneously satisfies all constraints: it respects companies right to conduct business by lowering the estimation cost and respects free speech rights by remaining statistically unbiased to a human ground truth. We encourage researchers in NLP and machine learning to broaden their purview beyond simply improving the performance of content moderation classifiers — a legal perspective opens up new academic venues, as we have found in our own work.

REFERENCES

[1] 2019. An Independent Report on How We Measure Content Moderation. https://about.fb.com/news/2019/05/dtag-report/

[2] Bobby Allyn. 2021. How One man's fight against an AOL troll sealed the tech industry's power. https://www.npr.org/2021/05/11/994395889/how-one-mans-fight-against-an-aol-troll-sealed-the-tech-industrys-power

[3] Javed A. Aslam, Virgil Pavlu, and Emine Yilmaz. 2006. A Statistical Method for System Evaluation Using Incomplete Judgments. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, Washington, USA) *(SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 541–548. https://doi.org/10.1145/1148170.1148263

[4] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, Munmun De Choudhury, Rumi Chunara, Aron Culotta, and Brooke Foucault Welles (Eds.). AAAI Press, 830–839. https://ojs.aaai.org/index.php/ICWSM/article/view/7347

[5] Paul N. Bennett and Vitor R. Carvalho. 2010. Online stratified sampling: evaluating classifiers at web-scale. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, Jimmy X. Huang, Nick Koudas, Gareth J. F. Jones, Xindong Wu, Kevyn Collins-Thompson, and Aijun An (Eds.). ACM, 1581–1584. https://doi.org/10.1145/1871437.1871677

[6] Howard S. Bloom. 1995. Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs. *Evaluation Review* 19, 5 (1995), 547–556. https://doi.org/10.1177/0193841X9501900504 arXiv:https://doi.org/10.1177/0193841X9501900504

[7] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 491–500. https://doi.org/10.1145/3308560.3317593

[8] Anu Bradford. 2012. The Brussels Effect. 107, 1 (2012), 1–68. https://heinonline.org/HOL/P?h=hein.journals/illlr107&i=1

[9] Valerie C. Brannon. 2019. Free speech and the regulation of social media content. *Congressional Research Service* 45650 (2019), 1–43.

[10] Arun Chaganty, Ashwin Paranjape, Percy Liang, and Christopher D. Manning. 2017. Importance sampling for unbiased on-demand evaluation of knowledge base population. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1038–1048. https://doi.org/10.18653/v1/D17-1109

[11] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy S. Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum. Comput. Interact.* 2, CSCW (2018), 32:1–32:25. https://doi.org/10.1145/3274301

[12] Philip M. Dixon, Aaron M. Ellison, and Nicholas J. Gotelli. 2005. Improving the precision of estimates of the frequency of rare events. *Ecology* 86, 5 (2005), 1114–1123. https://doi.org/10.1890/04-0601 arXiv:https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/04-0601

[13] Evelyn Douek. 2020. Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability. *SSRN Electronic Journal* (2020). https://doi.org/10.2139/ssrn.3679607

[14] William B. Fisch. 2002. Hate Speech in the Constitutional Law of the United States. *The American Journal of Comparative Law* 50 (2002), 463–492. https://doi.org/10.2307/840886 Publisher: American Society of Comparative Law.

[15] T. Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media.* Yale University Press. https://books.google.com.co/books?id=-RteDwAAQBAJ

[16] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234. https://doi.org/10.1177/2053951720943234 arXiv:https://doi.org/10.1177/2053951720943234

[17] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (Jan. 2020), 2053951719897945. https://doi.org/10.1177/2053951719897945 Publisher: SAGE Publications Ltd.

[18] Stefan Grundmann. 2011. »Inter-Instrumental-Interpretation«: Systembildung durch Auslegung im Europäischen Unionsrecht. *Rabels Zeitschrift für ausländisches und internationales Privatrecht / The Rabel Journal of Comparative and International Private Law* 75, 4 (2011), 882–932. https:

//www.jstor.org/stable/41304262 Publisher: Mohr Siebeck GmbH & Co. KG.

[19] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 386–396. https://doi.org/10.18653/v1/N18-1036

[20] Bruce Hoffman, Jacob Ware, and Ezra Shapiro. 2020. Assessing the Threat of Incel Violence. *Studies in Conflict & Terrorism* 43, 7 (2020), 565–587. https://doi.org/10.1080/1057610X.2020.1751459 arXiv:https://doi.org/10.1080/1057610X.2020.1751459

[21] Joel Kaplan. 2019. Improving enforcement and transparency of ads on Facebook. https://about.fb.com/news/2017/10/improving-enforcement-and-transparency/

[22] Müller Karsten. 2022. The effect of content moderation on online and offline hate. https://cepr.org/voxeu/columns/effect-content-moderation-online-and-offline-hate

[23] Katherine Keith and Brendan O’Connor. 2018. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4575–4585. https://doi.org/10.18653/v1/D18-1487

[24] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. All That’s Happening behind the Scenes: Putting the Spotlight on Volunteer Moderator Labor in Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (May 2022), 584–595. https://doi.org/10.1609/icwsm.v16i1.19317

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[26] David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical Obstacles to Deploying Active Learning. 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)*. Association for Computational Linguistics, Hong Kong, China, 21–30. https://doi.org/10.18653/v1/D19-1003

[27] Mark MacCarthy. 2019. A Consumer Protection Approach to Platform Content Moderation. *SSRN Electronic Journal* (2019). https://doi.org/10.2139/ssrn.3408459

[28] Sahar Massachi. 2022. How to save our social media by treating it like a city. https://www.technologyreview.com/2021/12/20/1042709/how-to-save-social-media-treat-it-like-a-city/

[29] Robinson Meyer. 2014. The primary way to report harassment on the Social Web is broken. https://www.theatlantic.com/technology/archive/2014/08/the-way-we-report-harassment-on-the-social-web-is-broken/378730/

[30] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=nzpLWnVAyah

[31] Brendan O’Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, William W. Cohen and Samuel Gosling (Eds.). The AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1536

[32] Penn State Department of Statistics. 2022. Lesson 6: Stratified sampling: Stat 506. https://online.stat.psu.edu/stat506/lesson/6

[33] High-Level Expert Group on AI. 2020. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment.* Publications Office. https://doi.org/10.2759/791819

[34] Art B. Owen. 2013. *Monte Carlo theory, methods and examples*.

[35] Joon Sung Park, Joseph Seering, and Michael S. Bernstein. 2022. Measuring the Prevalence of Anti-Social Behavior in Online Communities. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 451 (nov 2022), 29 pages. https://doi.org/10.1145/3555552

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[37] Neha Puri, Eric A. Coomes, Hourmazd Haghbayan, and Keith Gunaratne. 2020. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. *Human Vaccines & Immunotherapeutics* 16, 11 (2020), 2586–2593. https://doi.org/10.1080/21645515.2020.1780846 arXiv:https://doi.org/10.1080/21645515.2020.1780846 PMID: 32693678.

[38] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1668–1678. https://doi.org/10.18653/v1/P19-1163

[39] 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*. Association for Computational Linguistics, Valencia, Spain, 1–10. https://doi.org/10.18653/v1/W17-1101

[40] Burr Settles. 2012. *Active Learning*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00429ED1V01Y201207AIM018

[41] Alexandra A Siegel. 2020. Online hate speech. *Social media and democracy: The state of the field, prospects for reform* (2020), 56–88.

[42] Alexandra A. Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2021. Trumping Hate on Twitter? Online Hate Speech in the 2016 U.S. Election Campaign and its Aftermath. *Quarterly Journal of Political Science* 16, 1 (2021), 71–104. https://doi.org/10.1561/100.00019045

Prec. p	20%	10%	5%
0.1	865	3458	13830
0.059	1532	6127	24508
0.01	9508	38031	152122
0.001	95941	383762	1535047

Table 4. Number of samples needed for the random sampling estimator to estimate different prevalences, to different levels of precision. A normal approximation is assumed. The smaller the prevalence the more difficult it is to estimate precisely. Highlighted in bold is the prevalence of positives in the CivilComments dataset.

[43] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. *Proc. ACM Hum. Comput. Interact.* 3, CSCW (2019), 163:1–163:21. https://doi.org/10.1145/3359265

[44] James Vincent. 2020. Facebook bans holocaust denial content. https://www.theverge.com/2020/10/12/21512737/facebook-holocaust-denial-content-banned-hate-speech-policy-mark-zuckerberg

[45] Leijie Wang and Haiyi Zhu. 2022. How Are ML-Based Online Content Moderation Systems Actually Used? Studying Community Size, Local Activity, and Disparate Treatment. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 824–838. https://doi.org/10.1145/3531146.3533147

[46] Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2022. A Legal Approach to Hate Speech – Operationalizing the EU's Legal Framework against the Expression of Hatred as an NLP Task. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 53–64. https://aclanthology.org/2022.nllp-1.5

A POWER ANALYSIS FOR RANDOM SAMPLING

For reference, the number of samples to estimate different prevalences, to different levels of precision, with random sampling is listed in Table 4.

B TRAINING HYPERPARAMETERS FOR NLP MODELS

We experiment with a few supervised classification techniques. These classifier are used to score/rank the unlabeled examples, from which the strata are created.

Bag of words classifier. Using the TfidfVectorizer and LogisticRegression libraries from scikit learn [36], we trained a unigram bag-of-words classifier. In the preprocessing, sentences are tokenized by words, lowercased, and then the stopwords are dropped.

Roberta. Roberta [25] is a transformer-based language model which is pretrained on a large corpus of English data. Fine-tuning the model significantly improves classification tasks, and we find it to have strong performance for civilcomments as well. We use typical hyperparameters for small datasets [30], and make two adjustments that we found to be critical: increasing the batch size to 32 (we hypothesize that since the labels can be noisy, the gradients are unstable), and subsampling the negative examples so that the training data is balanced (the resulting training data has 11k examples).

Distilbert. Distilbert is trained with the same hyperparameters as Roberta.

C ANNOTATION DETAILS OF REDDIT COMMENTS

Preliminary analysis. Since the Digital Services Act does not specify to which ground truth "accuracy" can be measured, we explored a few options for academic study. For our study, our primary considerations are in choosing a class of speech that is 1) consistently annotated and 2) has a prevalence suitable for our annotation budget. With a small annotation budget, we can only estimate relatively large prevalences (1%-10%), but VLOPs are likely to have significantly larger annotation budgets and more accurate classifiers than us.

The subreddits we study here are chosen for their daily comment volume and prevalence of macro-norm violating comments [11, 35]. In our initial exploration, we used the Roberta civilcomments classifier from §4 to score comments in r/politics. We looked at within the highest scoring comments for categories of speech that seemed promising to estimate. We decided against using macro-norm violations, derived in Chandrasekharan et al. [11], because we found macro-norms such as “politically inflammatory” overly vague in annotation. In r/politics, a subreddit known for its toxicity, a large portion of comments came across as politically inflammatory. In the highest scoring comments by our toxicity classifier, we also almost never saw any content which was hate speech directed towards protected groups. We suspect that since r/politics is a heavily moderated subreddit, as with most popular subreddits, such egregious violations are almost always immediately removed.

Annotation. All annotation was conducted by the first author. The rate of annotation was about 600 examples per hour. In total 2400 samples were annotated (400 pilot samples per subreddit) totaling to about 6 hours of annotation time. During annotation, examples were given in random order, with the comment itself and the link to the comment on Reddit. The classifier score was hidden from the annotator to prevent the classification from biasing the annotator. Refer to Table 5 for a sample of our labels in personal attacks.

D ETHICAL CONSIDERATIONS

Automated decision-making systems in content moderation can have a negative impact on fundamental rights. Automated moderation decisions also raises ethical concerns about problematic training data or application of such systems perpetuating societal power dynamics [16]. We do not approve the use of these systems for censorship and have especially highlighted that automated content moderation needs to respect freedom of speech. We promote the use of stratified sampling: it involves humans and is unbiased with respect to a human defined ground truth. Finally, our work focuses on the *evaluation* of content moderation rather than directly making removal decisions, and we hope it can help further the study of fair interventions to reduce social media harms.

Our experiments are based on data sets that include postings qualifying as personal data in the sense of the EU General Data Protection Regulation. Even though the data sets have been made publicly available, our experiments qualify as "processing" and thus needs to be justified. We base this justification on Art. 6(1)(f), Art. 89 GDPR for purposes of scientific research.

Non top-level comment	Label
Your bias is showing lmao	⊕
Then join groups with people who are learning instead of trying to leech off players who want people on their level and a quick kill.	⊕
Instead of being an ignorant idiot when I've already answered you, why don't you just... scroll up? Like I said, I'm done here. I have a life... have the day you deserve.	⊕
At least you did attempt to think. That's great news!	⊕
I'll post whenever I feel like it. Deal with it	⊕
Never heard such a selfish statement in my life. It's not about just you, and your preference for raw sex puts your partner in danger.	⊕
Making these related is exactly what these assholes want. That's how terrorism fucking works. It's why ISIS takes credit for shit they didn't even do. So if y'all want to say that this was done because of a drag show, you're doing exactly what they want. They want the LGBTQ community to be blamed for this. Maybe it was, maybe it wasn't, but don't give PR for fucking terrorists and help victimize an already victimized group.	⊕
And? Who gives a fuck what people like? You asked what we thought about it	⊕
Thank you so much for answering! That is the compelling reason I would love to use! Would you say that it's possible for the judge to say yes if, like you said, I am fully supporting myself and providing for myself, and my parents provide me no support because I no longer need it? If I am in that position when I file, but one or both of my parents do not consent, will it be declined?	⊖
I mentioned it in the last session I had with them back in June when I also said I was looking at other trainers and they said they don't do refunds. But their policy says they only don't for owner training, not the board and train where they are the responsible party	⊖
oh got it, thanks! rn my situation here is that im preparing to file a case, however i was scared that the time it takes might be too long and such. right now i have a wide collection of recordings that ive read were admissible, and i was wondering how long the case should last if ever the crime is swiftly proven to have existed.	⊖
Do you know if that would be included in a standard copyright? Or commercial?	⊖
It's not really. Once you've contacted the police the decision to prosecute a crime becomes the DA & police decision. NAL, but I DO work for the Court in my state. (Not OPs state).	⊖
This isn't a very uncommon practice in older apartment buildings. My apartments in KY use the same system. It's called a Ratio Utility Billing System, but is usually just called RUBS. Heating and cooling split between all units.	⊖
New job it is then, what if there's quid pro quo happening? Which is what leads to them favoring the person's claims which leads to the firing?	⊖
Insurance will probably want a police report, so I would do that at the same time.	⊖

Table 5. A sample of annotations for personal attacks.