Operationalizing Content Moderation "Accuracy" in the Digital Services Act

Johnny Tian-Zheng Wei[1], Frederike Zufall[2,3], Robin Jia[1]

[1]University of Southern California
[2]Karlsruhe Institute of Technology
[3]Waseda University
jtwei@usc.edu, zufall@kit.edu, robinjia@usc.edu

Abstract

The Digital Services Act, recently adopted by the EU, requires social media platforms to report the "accuracy" of their automated content moderation systems. The colloquial term is vague, or open-textured—the literal accuracy (number of correct predictions divided by the total) is not suitable for problems with large class imbalance, and the ground truth and dataset to measure accuracy against is unspecified. Without further specification, the regulatory requirement allows for deficient reporting. In this interdisciplinary work, we operationalize "accuracy" reporting by refining legal concepts and relating them to technical implementation. We start by elucidating the legislative purpose of the Act to legally justify an interpretation of "accuracy" as precision and recall. These metrics remain informative in class imbalanced settings, and reflect the proportional balancing of Fundamental Rights of the EU Charter. We then focus on the estimation of recall, as its naive estimation can incur extremely high annotation costs and disproportionately interfere with the platform's right to conduct business. Through a simulation study, we show that recall can be efficiently estimated using stratified sampling with trained classifiers, and provide concrete recommendations for its application. Finally, we present a case study of recall reporting for a subset of Reddit under the Act. Based on the language in the Act, we identify a number of ways recall could be reported due to underspecification. We report on one possibility using our improved estimator, and discuss the implications and areas for further legal clarification.

1 Introduction

Many concerns have been raised about the role of social media in spreading harmful content such as online hate (Siegel 2020), and these concerns persist despite the fact that social media companies voluntarily moderate their content (Gillespie 2018).[1] On the other hand, content moderation raises the question of whether companies are interfering with users' freedom of speech. Research continues to demonstrate that voluntary moderation is problematic, and these issues will remain unresolved so long as social media companies have conflicting interests in prioritizing platform growth over providing safe online spaces (Pelley 2021).

[1]The full version of this paper including appendices can be found on arXiv at https://arxiv.org/abs/2305.09601.

In light of these conflicting interests, legislators have increasingly sought to enact legal requirements for content moderation. Regulating content moderation is a balancing act between conflicting rights: on one hand there is a need to protect the right to the freedom of speech, and on the other to curtail societal harms (Douek 2021a). This balance must be made all while considering the technical feasibility of such legal requirements, so as to respect companies' right to conduct business. Transparency reporting obligations, which require companies to release reports about their content moderation systems, are the first and necessary step in the regulatory landscape. Like any legal instrument, transparency requirements must respect the rights of all entities involved, and be defined with care to ensure their relevance.

The Digital Services Act, recently adopted by the EU, is a milestone legal instrument addressing content moderation. It includes a transparency obligation for social media platforms to report the "accuracy" of their content moderation systems. **However, the colloquial "accuracy" is underspecified, and the regulatory requirement, as it stands, allows for deficient reporting.** This is the departure point of our work: we operationalize "accuracy" by refining legal concepts and relating them to technical implementation. Our work provides an interpretation of content moderation "accuracy" in the Digital Services Act as precision and recall, which we show is both legally and technically sound. By providing this interpretation, we highlight areas for further legal clarification to ensure meaningful reporting. The argumentation here can be useful for future regulatory guidelines or legislative amendment.

Our interdisciplinary contribution can be viewed from two perspectives. From a legal perspective, we identify the sources of underspecification in the "accuracy" reporting requirement and show that precision and recall are appropriate instantiations which reflect the proportional balancing of Fundamental Rights of the EU Charter. From a technical perspective, we formulate the reporting of recall as unbiased statistical estimation and show how such estimation can be made efficient with stratified sampling. With the improved estimator, we demonstrate the feasibility of recall reporting in a case study on Reddit. Our paper contains three parts:

1. §3 starts with elucidating the legislative purpose of the Act to legally justify an interpretation of "accuracy" as precision and recall. These metrics remain informative

in class imbalanced settings and reflect the proportional balancing of the Fundamental Rights of the EU Charter.

2. §4 formulates the reporting of recall as an unbiased statistical estimation problem. Since its naive estimation can pose an disproportionate burden for companies, we propose the use of stratified sampling. In §5 we derive practical recommendations for applying stratified sampling through a simulation study.

3. §6 presents a case study of a recall reporting requirement for Reddit under the Digital Services Act. Based on the language in the Act, we identify a number of ways recall could be reported due to underspecification. We choose one possibility to report, and discuss the implications and areas for legal clarification.

2 Background and Related Work

Count-based metrics   At the time of writing, many social media companies voluntarily report basic content moderation statistics. Even without direct regulation, companies often provide transparency reports to build public trust in their platform.[2] For instance, Twitter reports a few count-based metrics such as the number of posts that Twitter took moderation action against in a given reporting period.[3] However, these basic statistics conflate multiple phenomena. When the number of takedowns of hate speech increases, it is not immediately clear whether it is due to an increase in hate speech, an increase in moderation, or a change in their content moderation policy.

Prevalence metrics   To the best of our knowledge, only Facebook, Youtube, and Snap publicly provide prevalence metrics. Facebook reports the prevalence of different categories of content violations over time (e.g. 2 in 1000 posts contained hate speech)[4]. As prevalence is a normalized measure (the number of content violations divided by the total amount of content on the platform) it is suitable to track increases in content violations. While their exact methods to compute prevalence are unclear, they note that "we calculate this metric by selecting a sample of content seen on Facebook and then labeling how much of it shouldn't be there."[5] Facebook's own internal studies deem prevalence as a major statistic and industry best practice (Plumb 2019). Research in social media has long used prevalence in online discussions to study the impact of real world events. For example, O'Connor et al. (2010) showed that presidential job approval can be well estimated by the prevalence of tweets mentioning the president which have positive sentiment.

Recall   There is a direct connection between prevalence and recall. By dividing the number of violations acted upon (i.e. the number of takedowns) against the total number of violations (the total number of posts multiplied by violation prevalence, see §4.2), we can calculate the recall of content moderation, or rate at which content violations are identified and acted against. Recall is a measure of moderation "thoroughness", and we speculate that these large companies may be computing a moderation recall internally to monitor moderation performance, by comparing prevalence with other statistics, such as number of takedowns (Liu 2019).

Metrics required by regulation   Almost all major social media platforms provide transparency reports for German users as specified by Germany's Network Enforcement Act (NetzDG).[6] NetzDG requires companies to periodically provide information and statistics related to the content moderation conducted at these companies (requirements which are succeeded by those in the Digital Services Act). Amongst the required information is also the community guidelines by which companies moderate content. These reports demonstrate the potential of regulation to standardize the set of transparency metrics reported on. By relating existing transparency requirements to content moderation recall, our work highlights how further legal clarification can ensure meaningful reporting. Recall could reflect general moderation effectiveness, and requiring it would encourage companies to periodically conduct studies of their platform design (Harling, Henesy, and Simmance 2023).

3 The Digital Services Act

The Digital Services Act (DSA), fully applicable since February 17th, 2024, introduces a comprehensive new framework for content moderation. Social media providers with more than 45 million average monthly active recipients in the Union (Art. 33(1) DSA), which currently includes U.S.-established platforms Twitter, Instagram, Youtube, 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.[7] This work focuses on the transparency reporting obligations on content moderation introduced by Art. 15(1).

Territorial scope   The DSA is a milestone legal instrument targeting content moderation. Hence, our study of the Act may have an impact beyond the EU—a phenomenon termed the "Brussels effect" (Bradford 2012). For one, nation-states outside of the EU will likely reference the DSA in considering legislation for their own jurisdictions. As companies may already have the infrastructure to satisfy EU requirements, adopting all or some of the DSA poses a low regulatory burden. Moreover, the DSA (like the GDPR) has a broad territorial scope, and requires companies operating in the EU to apply EU standards to all users, regardless of their nationality (Art. 2(1) DSA).

Considerations for future implementation   The Commission has begun to adopt implementing and designating

---

decisions for the DSA.[8] In addition, a new European Centre for Algorithmic Transparency has been established to support the Commission in analyzing transparency reports.[9] The effects of the Regulation will have to be reviewed by 2025 and 2027 (Art. 91 DSA)—allowing for an ongoing debate on its implementation and potential future amendments.

3.1 Scope of Content Moderation

Definition of "content moderation" The DSA defines "content moderation" as "activities [. . .] that are aimed, in particular, at detecting, identifying, and addressing" (Art. 3(t) DSA) the two following classes of content:

- *Illegal content.* This 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]

- *Content that violates terms and conditions.* 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). This includes the "community guidelines" as a part of the terms and conditions.

The DSA acknowledges a wide range of measures to "address" this identified content, including "measures taken that affect the availability, visibility, and accessibility of that illegal content or that information" (Art. 3(t) DSA).

Relation to Fundamental Rights The DSA, like any other regulatory instrument in EU law, is bound and to be interpreted and applied in accordance with the EU Charter of Fundamental Rights (CFR)[13]. While they only directly bind EU institutions (Art. 51 CFR), they also affect providers, as they guide the interpretation of provisions which regulate relations between citizens and private entities.[14] Namely,

Recital(3) of the DSA explicitly refers to three protected rights that are relevant here:

- *Freedom of expression and information (Art. 11 CFR)* protects online content from censorship and moderating measures. Moderating online posts interferes with this right and requires justification.

- *The right to non-discrimination (Art. 21 CFR)* forbids any discrimination based on grounds like sex, race, colour, ethnic or social origin and is especially reflected in the EU framework against the expression of hatred (Zufall et al. 2022).[15] To prevent discrimination, this right may provide the justification for the interference with freedom of expression.

- *Freedom to conduct a business (Art. 16 CFR)* is equally mentioned by the DSA and relevant for our context as any obligation imposed by the DSA to providers is interfering with this right. In addition to balancing out freedom of expression and the purpose of non-discrimination, the freedom to conduct a business also requires proportionality.

Balancing out these rights is a task not only performed by the legislator upon drafting regulation like the DSA, but also at the time of its subsequent executive implementation and judicial interpretation (Art. 52(5) CFR). Any automated content moderation system subject to the DSA that interferes with these rights requires a legal base in the form of law (Art. 52(1) CFR). This requirement reflects the democratic idea that any person who is subject to an interference must have the possibility to influence the origin of that interference through democratic elections. Here, the origin of interference is the content moderation system and the obligation to moderate content. Subsequently enacted laws by the legislation can again then build the legal foundation for limiting and balancing out these fundamental rights. Therefore, these rights continue to be relevant not only for regulatory implementing decisions, but also at the level of technical implementation.

3.2 "Accuracy" as a Transparency Requirement

Transparency requirements in the DSA not only are important legal instruments for revealing how content moderation systems work, but also for monitoring how they interfere with fundamental rights. Transparency requirements may also reflect the degree to which this balancing of rights has been carried out in favor of one right towards another. Therefore, transparency not only enables democratic control and judicial review, but also offers insight into the extent democratic legitimacy is carried through, to the final step of technical implementation.

Transparency requirements in the Act The DSA requires that Terms and Conditions must contain information on any policies, procedures, measures and tools used for the purpose of content moderation, including algorithmic

---

[8]https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

[9]https://algorithmic-transparency.ec.europa.eu/

[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. See also the recently adopted European Parliament Resolution of 18.01.2024, P9_TA(2024)0044.

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

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

[14]ECJ, C-360/10 – Sabam, 26.2.2012, para. 52.

[15]Usually, national constitutions contain a similar guarantee which is again reflected in national laws, e.g., Criminal Law, that once more constitute "illegal content" in the sense of the DSA.

decision-making and human review (Art. 14(1) DSA). Even further, providers are obliged to publish comprehensive reports at least once a year—in the case of VLOPs, every six months—on any content moderation they engaged in during the relevant period (Art. 15(1), Art. 42(1),(2) 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 that include their content moderation systems (Art. 34 DSA).

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".[16] From a technical perspective, the term "accuracy" raises a few questions: if we are estimating accuracy in its literal sense (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, so what metrics would be more appropriate? We identify "accuracy" as an *open-textured* term, which needs interpretation in novel contexts. As it stands, "accuracy" reporting required by the DSA will exhibit inconsistency, and social media providers may each make their own decisions regarding test data and evaluation metrics. They could choose to report the metric most beneficial for themselves, which would undercut the purpose of the regulation. By providing a legally and technically sound interpretation of "accuracy", our work highlights areas for further legal clarification to avoid an enforcement deficit.

What does "accuracy" refer to? Interpreting the notion of "accuracy" beyond its literal wording needs to be in line with the legislative purpose of the Act and requires justification. Future legal adaptations of our analysis depends on this legal justification. We attempt to elucidate and define the notion of "accuracy" in Art. 15(1)(e) of the DSA using a method of systematic interpretation in EU law: inter-instrumental interpretation (Grundmann 2011). 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 upcoming Artificial Intelligence Act (AI Act),[17] a parallel development in EU law related to specific high-risk AI systems. Similar to the DSA, the AI Act applies 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 AI systems, it outlines a general regulatory approach to automated classification systems.

With respect to "accuracy", Art. 15(1) of the AI Act 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." In this context, the AI Act (Art. 15(2)) calls upon the Commission to encourage "the development of benchmarks and measurement methodologies" for, inter alia, "the appropriate levels of accuracy". However, similarly to the DSA, the Act only mentions "accuracy" and not other metrics such as precision or recall. Upon further inspection, however, the legislative process reveals that the Act was inspired by the policy guidelines of the High Level Expert Group on AI (HLEG 2020). In these 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. Hence, any interpretation of "accuracy" beyond its literal wording in the DSA, should use an "appropriate metric depending on the application".

3.3 Metrics and Their Legal Implications

Having opened up "accuracy" to the possibility of broader interpretation, we consider its operationalization with basic accuracy metrics, and discuss their legal implications.

TP, FP, TN, and FN TP, FP, TN, and FN are the true positives, false positives, true negatives, and false negatives, respectively. This section discusses content moderation in terms of classifying illegal content (which is a perspective the DSA takes, see §3.1). In this context, TP and TN represent correctly identified illegal and free speech. Each FP is an individual violation to the freedom of expression (by misclassifying and acting against free speech), while each FN is a failure to identify and act against discriminatory content.

TP, FP, TN, and FN are primitives from which classification metrics are derived. Therefore, the reporting of TP, FP, TN, and FN, would be sufficient to calculate accuracy, precision, recall, F1, or any other metric. As FN and FP directly correspond to the interference of Fundamental Rights (listed in §3.1), we encourage regulators to reason about precision and recall, which are their normalized measures. Metrics such as accuracy or F1 (the harmonic mean of the precision and recall) can serve as useful summary statistics. However, only reporting summary statistics will obscure whether platforms are over- or undermoderating.

Accuracy This metric is by far mentioned the most in EU law, and defined as: Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$ which is the number of correct predictions divided by the total. Using accuracy as a metric for content moderation is problematic because of label-imbalance in the content moderation setting: in a huge set of content like social media posts, the illegal posts are usually far outnumbered by the posts that do not

violate specific laws. 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. In §5.1, we show that even a simple classifier can achieve an accuracy of 95% due to label imbalance in the data set. Alternatively, precision and recall are separate accuracies of the predicted positives and true positives, respectively, which remain informative in label imbalanced settings.

Precision Precision measures the fraction of examples positively predicted (e.g., identified by a content moderation system for removal) that are truly positive (e.g., illegal content, depending on the ground truth):

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{1}$$

As each FP is an individual violation to the freedom of expression (by misclassifying and acting against free speech), precision is a probabilistic reflection of the moderation system's respect to the right of freedom of expression. A low precision would indicate that much content which is not actually illegal is being positively identified and acted against. If a social media platform exceeds their legal obligation of content moderation, overmoderation would interfere with the right to freedom of expression. Content moderation is only justified if it does not present a disproportionate interference to this right, which is measured by precision.

Recall Finally, recall measures the fraction of truly positive content (e.g., illegal posts) that was predicted positive (e.g. identified by the moderation system for removal):

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{2}$$

As each FN is an individual failure to identify and act against discriminatory content, recall is a probabilistic reflection of the moderation system's respect to the right to non-discrimination. A low recall would indicate much illegal content is still visible on the platform. In this sense, recall well satisfies the purpose for which the DSA imposes a reporting obligation: to ensure compliance with EU or national laws against illegal content.

The precision-recall tradeoff The DSA has to balance the freedom of expression with the right to non-discrimination, and such a balance would be reflected mathematically by the precision-recall tradeoff.[18] A probabilistic classifier (i.e. one that predicts the probability of whether a given piece of content is illegal, and takes an action against it, such as removal, if the score is greater than a decision threshold) can trade off precision for recall and vica versa by tuning its decision threshold. By increasing the threshold, the classifier is more conservative in its positive predictions, which increases the precision at the cost of recall. Likewise, precision can be traded off for recall by decreasing the threshold. The decision threshold is an implicit but intentional setting set by the platforms that deploy these

---

[18]https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

classifiers, and requiring the reporting of precision and recall would make clear the effects of this implicit threshold (Douek 2021b). Regulators should recognize that, while improving both precision and recall simultaneously is nontrivial, trading them off is a technical option to strike the appropriate balance between fundamental rights.

4 Statistically Estimating Legal Requirements

Given that interpreting content moderation "accuracy" as precision and recall is legally sensible, we now discuss the technical implementation of precision and recall. This section starts with an example of statistical estimation intended for an interdisciplinary audience.

As an example, suppose a social media platform deploys an automated content moderation system that deletes content which violates their community guidelines. We may want to estimate the prevalence (or proportion) of content violations which remain visible on the platform. This prevalence would be the false negative rate of the deployed moderation system, and can be used to calculate the recall (more details in §4.2). We outline three components of statistical estimation here:

Expectation Statistical estimation begins with positing the ideal quantity we are estimating. Ideally, an annotator could label every single piece of visible content on the platform to determine whether it is violating or not. We could then directly calculate the false negative rate, which we denote as p (where $0 \leq p \leq 1$). Since p is derived from an exhaustive average, p is called an *expectation* in statistical terms. However, social media platforms accumulate millions of pieces of content a day, so exhaustive annotation would not be feasible. Therefore, the true p will be unknown.

Sampling The only practical alternative is to estimate the false negative rate p by *sampling* some visible content and then annotating this sample. The prevalence of content in the sample, which we denote as $\hat{p}$, can be used as an estimate for p. The number of samples directly relates to the implementation cost: each sample needs to be labelled by an annotator.

Standard error A better estimate $\hat{p}$ is one that is close to p, on average. The average distance between $\hat{p}$ and p is a *standard error* in statistical terms. If the estimate $\hat{p}$ is unbiased, more samples results in smaller standard error. Estimation error is then a function of these two components:

- **Unbiasedness** For legal reporting, we should provide a *statistically unbiased* estimate $\hat{p}$ which, on average, reflects the false negative rate p. When estimating $\hat{p}$, two details ensure statistical unbiasedness. The first is using a "representative" sampling method like random sampling. Stratified sampling is also unbiased by emulating a "representative" sample (see §4.3). Second, human ground truth annotations are required. Solely relying on classifiers introduces bias because they make errors relative to the ground truth (Gorwa, Binns, and Katzenbach 2020).

- **Number of samples (cost)** An unduly burdensome reporting requirement would interfere disproportionately with the right to conduct business (Art. 16 CFR). Thus,

our statistical analyses focuses on cost: *how many* annotations are needed to provide a reasonably good estimate? In general, a larger sample yields better estimates. On the other hand, as p approaches 0, it requires more samples to accurately estimate e.g. if $p = 0.02$, 10 samples would likely yield estimates $\hat{p} = 0.0$ or $\hat{p} = 0.1$, where the sample contains 0 or 1 pieces of violating content. More samples would be needed to refine the granularity of $\hat{p}$.

4.1 Estimating Precision

Auditing content moderation precision is qualitatively difficult but statistically simple, and so is not the focus of our work. In short, precision can be well estimated by annotating a random sample of positively predicted content.

4.2 Estimating Recall

Measuring recall poses a challenging statistical problem, which is the focus of our work. To compute recall, both TP and FN are needed. TP can be computed from the precision (the positives multiplied by the precision), but estimating FN requires looking through the predicted negatives of the content moderation system (i.e. visible content on the platform during the last reporting period) to determine the number of false negatives. Similarly, recall is challenging to measure for tasks like information retrieval, as the false negatives are the relevant webpages missed by the retrieval system across the web (Aslam, Pavlu, and Yilmaz 2006).

Typically, the prevalence of misclassified violations amongst all content is rare, e.g. 1 in 1000, so estimating the false negative rate requires a large number of samples to provide a reasonable estimate. The simplest way to estimate p is to apply the random sampling estimator, where a random sample of negatives is collected and annotated to compute $\hat{p}$ as an estimate for p. When estimating a small value like p, we want the estimation error to be small, relative to p. In other words, we are estimating the rare probability p within some percentage accuracy. In statistical terms, we want to ensure the coefficient of variance $\text{CV}_{\hat{p}} = \text{SE}_{\hat{p}}/\hat{p}$ is low i.e. $\text{CV}_{\hat{p}} \leq 20\%$, where $\text{SE}_{\hat{p}}$ is the standard error of estimate $\hat{p}$ (Dixon, Ellison, and Gotelli 2005).

4.3 Stratified Sampling

To reduce the burden of recall reporting, we propose using stratified sampling instead of random sampling. Stratified sampling can provide a better estimate with the same number of samples (Owen 2013). Crucially, it is *unbiased* and *more efficient* than a random sampling estimator. The procedure is as follows: a learned classifier is first used to partition e.g. the visible comments into bins. Then, the estimator is computed by annotating a random sample within each stratum, and combining the prevalences across stratum in an unbiased manner.

Intuition A good stratification can significantly increase the efficiency of the stratified sampling estimator over random sampling. If some strata have a very low or very high prevalence of violating comments, few samples are needed in those strata to estimate the prevalences well. This allows us to allocate our annotation effort on estimating strata with

medium prevalences (which have higher variance). Stratified sampling is unbiased even if the stratification classifier has undesirable properties such as low accuracy or low fairness (Borkan et al. 2019), as a human is the final decision maker. A suboptimal classifier may cause the estimator to be less efficient, but it cannot introduce statistical bias.

4.4 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 provide technical details for each choice.

Binning An effective way to create strata is with a binning classifier trained to predict the probability that a given piece of content violates community guidelines. With the predictions scores from the binning classifier, content can be binned in a few ways by their score. The binning classifiers gives each comment a score from $[0, 1]$. The number of bins L is a hyperparameter, which we investigate in §5. **(1)** *Equal width binning*. 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 binning 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, an optimal sample allocation minimizes the variance of the estimator. 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 (2010). 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 psuedo-counts (adding 1 positive and 1 negative example) to ensure each stratum is allocated samples.

4.5 Analyzing the Burden of Estimation

The statistical analyses here are focused on cost i.e. how many annotations are needed to provide a good estimate? The number of annotations needed to achieve a fixed estimation error can be calculated analytically for both random and stratified sampling estimators. A few technical details are provided here and we defer the rest to the appendix.

We consider the goal to report $\hat{p}$ to within 20% of p, with 95% probability. In statistical terms, we calculate the number of samples required for the 95% confidence interval of $\hat{p}$ to be a fixed width relative to p.[19] The coefficient of variance is effectively required to be $\mathrm{CV}_{req} = 0.2/z_{.95}$, and in turn the standard error is required to be $\mathrm{SE}_{req} = (0.2)\hat{p}/z_{.95}$, where $z_{(1-\alpha)}$ is test statistic of the normal variable. Since SE_{req} is a function of the total number of annotations collected, we can solve for this number for each estimator.

5 Efficiently Estimating Recall for a Filter

In this section, we determine which choices of binning, allocation (described in §4.4), and classifier are best for estimating recall with stratified sampling. The efficient and robust estimator derived in this section is used to estimate a content moderation recall for Reddit in §6, where we will continue to provide legal analysis on areas of underspecification in the "accuracy" reporting requirement.

We use CivilComments (Borkan et al. 2019) as our testbed and emulate a content moderation system by constructing a keyword-based comment filter. While Civilcomments is annotated for toxicity and is not an operationalized legal definition or community guideline, this setting provides a large dataset to study recall estimation on a related task. Similar to pool-based active learning (Settles 2012), we assume in our simulated experiments that the dataset is initially unlabeled, and we "annotate" whether an example is a false negative by revealing the label from the dataset.

5.1 A Simple Keyword Filter

To emulate an automated content moderation system, we construct a keyword-based filter which removes comments in the Civilcomments dataset (see §5.2). We choose the list of keywords using the top 10 positive features from an unigram logistic regression trained to predict comment toxicity.

In total, the dataset contains about 100K toxic comments out of 1.7M comments, and the overall prevalence of toxicity is 5.9%. The filter achieves a precision of 58%, removing 57K comments with 32K of these actually being toxic. Amongst the unfiltered examples, the prevalence is 4.1%. The filter would only achieve a recall of 33%, removing about 33K of the total 100K toxic comments. However, the accuracy would be nearly 95%.

---

[19]Once we have a statistical estimate $\hat{p}$ for the false negative rate p, the confidence intervals on $\hat{p}$ can propagate to a confidence interval on recall. In the case where we know the exact number of true positives TP, the confidence intervals on p can directly be translated to confidence intervals on FN, and then to recall, by plugging in the upper and lower estimates of p into the recall equation. If TP is not a known quantity and also an estimate, confidence intervals can be obtained through bootstrap simulation.

Note that the high accuracy obscures the fact that only 33% of the toxic comments were found and removed. In our simulated experiments, the goal is to statistically estimate the recall using as few labeled examples as possible. By estimating the prevalence of toxicity in the unfiltered comments, we can calculate recall. In our experiments, we compare costs when estimating the prevalence of 4.1% to within 20%. If prevalence is estimated as such, the confidence intervals on the recall would translate to $[28.9\%, 37.9\%]$.

5.2 Experimental Setup

Dataset The CivilComments dataset (Borkan et al. 2019) contains comments labeled by their toxicity, from the archive of the Civil Comments platform, 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 100K random examples of the training set for training, and the rest of the 1.7M training examples for simulation. This creates a large testing setting which better represents content moderation, where the number of examples to train the model is small but it is used for a large number of predictions. This dataset naturally presents a rare prevalence estimation problem as 5.9% of comments are toxic.

Binning classifier Roberta (Liu et al. 2019) is a transformer-based language model pretrained on a large corpus of English data, which has strong performance for CivilComments after finetuning. We use typical hyperparameters for small datasets (Mosbach, Andriushchenko, and Klakow 2021), and make two critical adjustments: 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).

5.3 Discussion

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.

Pilot sampling under different prevalences We examine the robustness of pilot sampling to different class imbalances in Figure 2. For each example, civilcomments has a real valued toxicity label 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 the setting studied in Figure 1. In Figure 2 we present results for different prevalences by varying the threshold among $\{0.48, 0.83, 0.95\}$. Stratified sampling provides significant efficiency over random sampling. The number of pi-
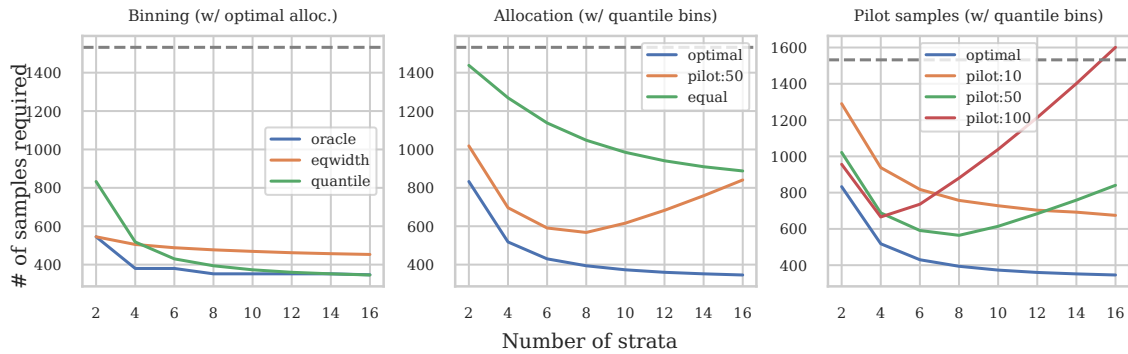
Figure 1: Number of samples required to estimate the prevalence of toxicity (4.1%) within 20% in the unfiltered CivilComments, with different number of strata. Binning is based on predicted scores from a finetuned Roberta model. All pilot results are averaged over 30 trials. Dotted line represents random sampling. 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.

lot samples 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 fewer samples are available for each bin to estimate the standard deviation with.

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.

6 Recall Reporting in Practice

In this section, we present a case study of recall reporting for Reddit. As the requirement currently underspecifies the technical implementation, we identify a number of ways recall could be reported. We then choose one to report and discuss the implications and areas for legal clarification. We choose to focus on Reddit for two reasons: first, Reddit qualifies as an "online platform" as defined in Art. 3(i) DSA and therefore is obliged to provide transparency reports annu-

ally (Art. 15(1) DSA).[20] Second, Reddit data is most readily available for academic study.

6.1 Addressing Underspecification

In implementing a technical evaluation to satisfy "accuracy" reporting for the DSA, three basic specifications are needed:

Evaluation metric As discussed in §3, we establish recall as a legally sensible metric, which we showed in §5 can be efficiently estimated with stratified sampling.

Ground truth The DSA addresses two classes of content (see §3.1). To report recall, we must consider how the true positives and true negatives can be annotated:

- *Illegal content.* To identify illegal content, legal definitions (e.g., punishable hate speech) would need to be operationalized as clear annotation guidelines (Zufall et al. 2022). To be feasible at the scale social media requires, such identification likely needs to be possible for a skilled person with no legal training.

- *Content that violates terms and conditions.* A convenience in identifying whether speech violates community guidelines as a part of terms and conditions 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 transparency measures about the guidelines themselves (see Art. 14 DSA).

As the scope of the DSA applies to all content, and all platforms have community guidelines (Gillespie 2018), this specification may be general enough to allow for explicit legal clarification. However, regulators should take caution when specifying recall on illegal content. Such reporting is only feasible if the respective legal bases defining content illegality are amenable to operationalization.

---

[20]https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413
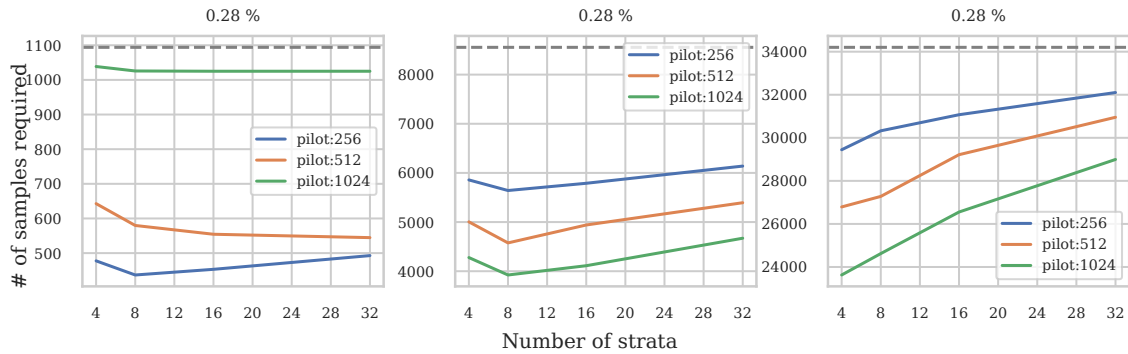
Figure 2: Number of samples required to estimate the prevalence of toxicity within 20%, for different prevalences. All pilot results are averaged over 100 trials. The numbers associated with the pilot is the *total* number of pilot samples across all bins. By choosing appropriate powers of two, each bin always has an integer number of pilot samples. Dotted lined represents random sampling. 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.

Test set Finally, the set of content (or test set) to report the accuracy metric over needs to be decided. Since the DSA specifies bi-annual or annual reporting periods, it would be natural to evaluate on content accumulated within the last reporting period. The DSA does not provide further guidance, and we speculate two appropriate test sets:

- *All visible content.* An accuracy metric could be evaluated over all visible content on the platform, which would also follow industry best practices on how prevalence metrics are reported (see §2).

- *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. When a notice is submitted, providers are liable if they do not act expeditiously to provide a legal examination and remove or disable access to this content if it is illegal (Art. 16(3), Art. 6(1)(b) DSA). As platforms are expected to review this content, the moderation recall of this subset may be especially important. However, anyone can flag content without concrete reason, possibly even with malicious intent (Meyer 2014).

The specification of a test set will be heavily platform dependent. For Reddit, content may mean posts or comments, but for Youtube it may mean videos. In Snapchat, most content is not publicly visible, and it may be more appropriate to calculate recall over views. Regulators should be wary of overgeneralizing, as social media platforms have diverse functionality (Gillespie et al. 2020). Underspecifying the test set allows for custom implementation and may encourage meaningful reporting.

Our study We study the moderation recall of subreddits, in terms of content removal as the means of moderation.[21] For the purposes of our study, we choose to report the recall

of personal attacks on all visible comments. We find this to reasonably satisfy the requirement as Reddit has community guidelines against such speech and most of the discussion happens in the comments. Practically, the other options require data or annotations that are not readily available for academic study. In our preliminary analysis of comments, we found that personal attacks can be identified consistently. This is corroborated by Habernal et al. (2018), which reports that personal attacks in r/changemyview have high inter-annotator agreement.

6.2 Experimental Setup

Study of interest We restrict our study to a recent dump of Reddit (December 2022) on Pushshift.io (Baumgartner et al. 2020). We select 6 subreddits with a range of macro-norm violation prevalence, as estimated in Park, Seering, and Bernstein (2022), and that have more than 1000 comments per day[22]:

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

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.

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[23] 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.

---

[21]Content removal is the primary means of moderation here, see https://support.reddithelp.com/hc/en-us/articles/15484284113172-What-mods-can-do

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

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

	$\hat{p}_{st}$	95% CI	n for $\text{CV}_{\hat{p}_{st}} \leq 20\%$		Reduction w/ strat.	Comments removed	Mod. recall	95% CI
			Stratified	Random				
r/politics	7.00%	[6.42%, 7.58%]	1033	1401	26%	0.77%	9.91%	[9.22%, 10.71%]
r/AskReddit	4.75%	[4.29%, 5.21%]	1425	2142	33%	0.70%	12.92%	[11.91%, 14.11%]
r/sex	3.00%	[2.63%, 3.37%]	2477	3556	30%	2.06%	40.72%	[37.94%, 43.93%]
r/pcmasterrace	4.00%	[3.61%, 4.39%]	1488	2567	42%	0.89%	18.20%	[16.86%, 19.78%]
r/wow	4.25%	[3.76%, 4.74%]	2078	2455	15%	0.75%	14.97%	[13.63%, 16.60%]
r/legaladvice	1.50%	[1.25%, 1.75%]	5027	7591	33%	7.65%	83.62%	[81.38%, 85.97%]

Table 1: 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 and efficiency for estimation with random and pilot stratified sampling is provided assuming the lower bound of prevalence.

Removed comments While publicly viewable posts from Reddit are accessible through the Pushshift data stores, records for removed comments are incomplete. Therefore, we can only derive an upper bound of content moderation recall by assuming every removal decision is a true positive.

Personal attacks Using the data from Habernal et al. (2018), we finetuned a Roberta binning classifier on a balanced training set with an equal number of personal attacks and negatives from r/changemyview. We label comments as personal attacks according to Habernal et al. (2018).

6.3 Discussion

Results Refer to Table 1. As a sanity check, we see that ranking subreddits by personal attack prevalence aligns closely with their ranking by macro-norm violation as estimated in Park, Seering, and Bernstein (2022). By the prevalences, we can conclusively determine that r/politics has the biggest problem with personal attacks out of all the subreddits we examine. By moderation recall, r/politics is also low. While they have a rule against personal attacks, we speculate that moderators avoid removing too many comments to appear politically neutral. For r/legaladvice, personal attack prevalence is already low, and moderators also remove many comments that are off topic, which may inflate the true positives and consequently the recall. In terms of methodology, applying our version of stratified sampling often yields substantial sample reductions. These gains are present despite distribution shift, where the binning classifier was trained on r/changemyview but tested on other subreddits.

Confidence intervals Even though we collected a small number of samples, the confidence intervals on the recall are narrow because the relative difference between the true positives and false negatives was often large. This means that as the number of personal attacks removed surpasses the prevalence, the recall becomes easier to estimate. Since we were able to provide tight bounds around the recall with a small number of annotator hours, we expect it feasible for Reddit to report more comprehensive recall measures annually.

Discrepancies with the reporting obligation There are two discrepancies between our experiment and the actual reporting obligation of the DSA. First, our recall is reported over a month while Reddit only needs to provide transparency reports every year. Our method can easily be adopted to a larger pool and the estimation difficulty only increases in the larger pool if the prevalence decreases. Second, the removed comments we consider could have been removed through either automated or manual means, while only recall reporting of automated tools are required by the DSA. Since we do not have records of how the comments were removed, we cannot easily provide an isolated accuracy of the automated systems, but our method is general. However, we point out that the recall presented is in line with current industry prevalence metrics, and we additionally suggest legislators could expand the scope of the requirement to an aggregate moderation accuracy.

7 Conclusion

In this work, we identified "accuracy" in the DSA as an open-textured term, which requires further interpretation. To operationalize the requirement, we considered the regulatory purposes standing behind the DSA and the underlying fundamental rights which are affected. The DSA balances freedom of expression with the right to non-discrimination, and we find precision and recall to be an appropriate instantiation of these rights, where balancing is reflected in the precision-recall tradeoff. We derive an efficient and robust estimator, which we use to compute a moderation recall for Reddit. By providing a legally and technically sound interpretation of "accuracy", we highlight the implications and areas for further legal clarification. In light of the insights here, we hope regulators can provide further legal or regulatory clarification to prevent enforcement deficits and better fulfill the regulatory purpose of the DSA.

8 Ethical Considerations

Risks of content moderation Content moderation can violate the right to the freedom of expression. Automated content moderation further raises ethical concerns about problematic training data or application of such systems perpetuating societal power dynamics (Gillespie 2020). 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.

Risks of regulatory capture Regulation can be co-opted to serve minority interests. Facebook, for instance, has openly requested for additional social media regulation, which is speculated to serve their interest in preventing competition (Macias 2020). Our position on regulation is not neutral either, as this work is rooted in existing regulation. However, the entirety of this work is dedicated to content moderation recall's legal and technical sensibility. Our proposals introduce a meaningful transparency requirement which introduces strictly more reporting burden. A large portion of this work is dedicated to technical feasibility, which is especially important in considering the compliance of new companies and preventing regulatory capture.

Data statement 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 before, 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.

References

Aslam, J. A.; Pavlu, V.; and Yilmaz, E. 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*, SIGIR '06, 541–548. New York, NY, USA: Association for Computing Machinery. ISBN 1595933697.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. In Choudhury, M. D.; Chunara, R.; Culotta, A.; and Welles, B. F., eds., *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*, 830–839. AAAI Press.

Bennett, P. N.; and Carvalho, V. R. 2010. Online stratified sampling: evaluating classifiers at web-scale. In Huang, J. X.; Koudas, N.; Jones, G. J. F.; Wu, X.; Collins-Thompson, K.; and An, A., eds., *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, 1581–1584. ACM.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In Amer-Yahia, S.; Mahdian, M.; Goel, A.; Houben, G.; Lerman, K.; McAuley, J. J.; Baeza-Yates, R.; and Zia, L., eds., *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, 491–500. ACM.

Bradford, A. 2012. The Brussels Effect. 107(1): 1–68.

Dixon, P. M.; Ellison, A. M.; and Gotelli, N. J. 2005. Improving the precision of estimates of the frequency of rare events. *Ecology*, 86(5): 1114–1123.

Douek, E. 2021a. Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability. *Columbia Law Review*, 121(3): 759–834.

Douek, E. 2021b. Governing online speech: From" posts-as-trumps" to proportionality and probability. *Colum. L. Rev.*, 121: 759.

Gillespie, T. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press. ISBN 9780300173130.

Gillespie, T. 2020. Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2): 2053951720943234.

Gillespie, T.; Aufderheide, P.; Carmi, E.; Gerrard, Y.; Gorwa, R.; Matamoros-Fernández, A.; Roberts, S. T.; Sinnreich, A.; and Myers West, S. 2020. Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4).

Gorwa, R.; Binns, R.; and Katzenbach, C. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1): 2053951719897945. Publisher: SAGE Publications Ltd.

Grundmann, S. 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): 882–932. Publisher: Mohr Siebeck GmbH & Co. KG.

Habernal, I.; Wachsmuth, H.; Gurevych, I.; and Stein, B. 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)*, 386–396. New Orleans, Louisiana: Association for Computational Linguistics.

Harling, A.-S.; Henesy, D.; and Simmance, E. 2023. Transparency Reporting: The UK Regulatory Perspective. *Journal of Online Trust and Safety*, 1(5).

HLEG, H.-L. E. G. o. A. 2020. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office.

Liu, Y. 2019. Estimating the prevalence of rare events — theory and practice.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V.

2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Macias, A. M. 2020. Facebook CEO Mark Zuckerberg calls for more regulation of online content.

Meyer, R. 2014. The primary way to report harassment on the Social Web is broken.

Mosbach, M.; Andriushchenko, M.; and Klakow, D. 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.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In Cohen, W. W.; and Gosling, S., eds., *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*. The AAAI Press.

Owen, A. B. 2013. *Monte Carlo theory, methods and examples*.

Park, J. S.; Seering, J.; and Bernstein, M. S. 2022. Measuring the Prevalence of Anti-Social Behavior in Online Communities. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Pelley, S. 2021. Whistleblower: Facebook is misleading the public on progress against hate speech, violence, misinformation.

Plumb, R. I. 2019. Exploring feedback from data and governance experts: A research-based response to the Data Transparency Advisory Group report.

Settles, B. 2012. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

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

Zufall, F.; Hamacher, M.; Kloppenborg, K.; and Zesch, T. 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*, 53–64. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.