TAR on Social Media: A Framework for Online Content Moderation

Eugene Yang[1], David D. Lewis[2] and Ophir Frieder[3]

[1]*IR Lab, Georgetown University, Washington, DC, USA*
[2]*Reveal Brainspace, Chicago, IL, USA*
[3]*IR Lab, Georgetown University, Washington, DC, USA*

Abstract

Content moderation (removing or limiting the distribution of posts based on their contents) is one tool social networks use to fight problems such as harassment and disinformation. Manually screening all content is usually impractical given the scale of social media data, and the need for nuanced human interpretations makes fully automated approaches infeasible. We consider content moderation from the perspective of technology-assisted review (TAR): a human-in-the-loop active learning approach developed for high recall retrieval problems in civil litigation and other fields. We show how TAR workflows, and a TAR cost model, can be adapted to the content moderation problem. We then demonstrate on two publicly available content moderation data sets that a TAR workflow can reduce moderation costs by 20% to 55% across a variety of conditions.

Keywords

Technology-assised review, active learning, social media, content moderation, cost analysis

1. Introduction

Online social networks are powerful platforms for personal communication, community building, and free expression. Unfortunately, they can also be powerful platforms for harassment, disinformation, and perpetration of criminal and terrorist activities. Organizations hosting social networks, such as Facebook, Twitter, Reddit, and others, have deployed a range of techniques to counteract these threats and maintain a safe and respectful environment for their users.

One such approach is content moderation: removal (hard moderation) or demoting (soft moderation) of policy-violating posts [1, 2]. Despite recent progress in machine learning, online content moderation still heavily relies on human reviews [3]. Facebook's CEO Mark Zuckerberg stated that language nuances could get lost when relying on automated detection approaches, emphasizing the necessities for human judgments. [1] Ongoing changes in what is considered inappropriate content complicates the use of machine learning [4]. Policy experts have argued that complete automation of content moderation is socially undesirable regardless of algorithmic accuracy [5].

It is thus widely believed that both human moderation and automated classification will be required for online content moderation for the foreseeable future [1, 5, 6]. This has meant not just capital investments in machine learning tools for moderation, but also massive ongoing personnel expenses for teams of human reviewers [7].

Surprisingly, the challenge of reducing costs when both machine learning and manual review are necessary has been an active area of interest for almost two decades, but in a completely different area: civil litigation. Electronic discovery (eDiscovery) projects involve teams of attorneys, sometimes billing the equivalent of hundreds of euros per person-hour, seeking to find documents responsive to a legal matter [8]. As the volume of electronically produced documents grew, machine learning began to be integrated in eDiscovery workflows in the early 2000s, a history we review elsewhere [9].

The result in the legal world has been *technology-assisted review* (TAR): human-in-the-loop active learning workflows that prioritize the most important documents for review [10, 11]. One-phase (continuous model refinement) and two-phase (with separate training and deployment phases) TAR workflows are both in use [9, 12].

Because of the need to find most or all relevant documents, eDiscovery has been referred to as a *high recall review* (HRR) problem [13, 14, 15]. HRR problems also arise in systematic reviews in medicine, sunshine law requests, and other tasks [16, 17, 18]. Online content moderation is an HRR problem as well, in that a very high proportion of inappropriate content should be identified and removed.

Our contributions in this paper are two-fold. First, we describe how to adapt TAR and its cost-based evaluation framework to the content moderation problem. Second,

---

[1]https://www.businessinsider.com/zuckerberg-nuances-content-moderation-ai-misinformation-hearing-2021-3

we test this approach using two publicly available content moderation datasets. Our experiments show substantial cost reductions using the proposed TAR framework over both manual review of unprioritized documents and training of prioritized models on random samples.

2. Background

Content moderation on online platforms is a necessity [19, 20] and has been argued by some to be the defining feature of an online platform [6]. Despite terms of service and community rules on each platform, users produce inappropriate content, particularly when anonymous [21]. Inappropriate content includes toxic content such as hate speech [22], offensive content [23], and mis / disinformation [4, 23]. It also includes content that is inappropriate for legal or commercial reasons, such as potential copyright violations [5, 24].

The identification of toxic content can require subtle human insight [4, 22], both due to attempts at obfuscation by posters, and because the inappropriateness of the content is often tied to its cultural, regional, and temporal context [1, 3]. Mis- and disinformation often consists of subtle mixtures of truthful and misleading content that require human common sense inferences and other background knowledge [4, 23].

Social media organizations have deployed numerous techniques for implementing community policies, including graph- and time-based analyses of communication patterns, user profile information, and others [25]. Our focus here, however, is on methods that use the content of a post.

Content monitoring falls into three categories: manual moderation, text classification, and human-in-the-loop methods. The latter two approaches leverage machine learning models and are sometimes collectively referred to as *algorithmic content moderation* in policy research [5].

Manual moderation is the oldest approach, dating back to email mailing lists. It is, however, extremely expensive at the scale of large social networks and suffers potential human biases. Additionally, mental health concerns are an issue for moderators exposed to large volumes of toxic content [25, 26, 27].

The simplest text classification approaches are keyword filters, but these are susceptible to embarrassing mistakes[2] and countermeasures by content creators. More effective text classification approaches to content moderation are based on supervised machine learning [28, 29]. Content types that have been addressed include cyberbullying [29, 30, 31, 32], hate speech [22, 31, 33, 34, 35, 36] or offensive language in general [23, 37, 38, 39, 40, 41, 42].

However, some moderation judgments are inevitably too subtle for purely automated methods[3], particularly when content is generated with the intent of fooling automated systems [1, 25, 43]. Content that is recontextualized from the original problematic context, for example, through reposting, screenshotting, and embedding in new contexts complicates moderation [2]. Additionally, bias in automated systems can also arise both by learning from biased labels and from numerous other choices in data preparation and algorithmic settings [27, 44, 45]. Biased models risk further marginalizing and disproportionately censoring groups that already face discrimination [1]. Differences in cultural and regulatory contexts further complicate the definition of appropriateness, creating another dimension of complexity when deploying automated content moderation [4].

Human-in-the-loop approaches, where AI systems actively manage which materials are brought to the attention of human moderators, attempt to address the weaknesses of both approaches while gathering training data to support supervised learning components [25, 46]. Filtering mechanisms that proactively present only approved content (pre-moderation) and/or removal mechanisms that passively take down inappropriate ones are used by platforms depending on the intensity [4]. Reviewing protocols could shift from one to the other based on the frequency of violations or during a specific event, such as elections[4]. Regardless of the workflows, the core and arguably the most critical components is reviews. However, the primary research focus of human-in-the-loop content moderation has been on classification algorithm design and bias mitigation, rarely on the investigation of the overall workflow.

Like content moderation, eDiscovery is a high recall retrieval task applied to large bodies of primarily textual content (typically enterprise documents, email, and chat) [11, 12]. Both fixed data set and streaming task structures have been explored, though the streaming context tends to bursty (e.g., all data from a single person arriving at once) rather than continuous. Since cost minimization is a primary rationale for TAR [47], research on TAR has focused on training regimens and workflows for minimizing the number, or more generally the cost, of documents reviewed [9, 12]. A new TAR approach is typically evaluated for its ability to meet an effectiveness target while minimizing cost or a cost target while maximizing effectiveness [18, 48, 49]. This makes approaches developed for TAR natural to consider for content moderation.

---

[2]https://www.techdirt.com/articles/20200912/11133045288/paypal-blocks-purchases-tardigrade-merchandise-potentially-violating-us-sanctions-laws.shtml

[3]https://venturebeat.com/2020/05/23/ai-proves-its-a-poor-substitute-for-human-content-checkers-during-lockdown/

[4]https://www.washingtonpost.com/technology/2020/11/07/facebook-groups-election/

3. Applying TAR to Content Moderation

In most TAR applications, at least a few documents of the (usually rare) category of interest are available at the start of the workflow. These are used to initialize an iterative pool-based active learning workflow [50]. Reviewed documents are used to train a predictive model, which in turn is used to select further documents based on predicted relevance [51], uncertainty [52], or composite factors. Workflows may be batch-oriented (mimicking pre-machine learning manual workflows common in the law) or a stream of documents may be presented through an interactive interface with training done in the background. These active learning workflows have almost completely displaced training from random examples when supervised learning is used in eDiscovery.

Two workflow styles can be distinguished [9]. In a *one-phase workflow*, iterative review and training simply continues until a stopping rule is triggered [49, 53, 54]. Stopping may be conditioned on estimated effectiveness (usually recall), cost limits, and other factors [53, 55, 56]. *Two-phase workflows* stop training before review is finished, and deploy the final trained classifier to rank the remaining documents for review. The reviewed documents are typically drawn from the top of the ranking, with the depth in the ranking chosen so that an estimated effectiveness target is reached [18, 48]. Two-phase workflows are favored when labeling of training data needs to be done by more expensive personnel than are necessary for routine review.

The cost of both one- and two-phase TAR workflows can be captured by in a common cost model [9]. The model defines the total cost of a one-phase review terminated at a particular point as the cost incurred in reviewing documents to that point, plus a penalty if the desired effectiveness target (e.g., a minimum recall value) has not been met. The penalty is simply the cost of continuing on to an optimal second-phase review from that point, i.e. the minimum number of prioritized documents is reviewed to hit the effectiveness target. For a two-phase workflow, we similarly define total cost to be the cost of the training phase plus the cost of an optimal second phase using the final trained model.

These costs in both cases are idealizations in that there may be additional cost (e.g. a labeled random sample) to choose a phase two cutoff citecikmpaper. However, the model allows a wide range of workflows to be compared on a common basis, as well as allowing differential costs for review of positive vs. negative documents, or phase one vs. phase two documents.

While developed for eDiscovery, the above cost model is also a good fit for content moderation. As discussed in the previous section, the human-in-the-loop modera-

shut up mind your own business and go f*** some one else over

(a) Wikipedia collection.

: being in love with a girl you dont even know yours is sadder
: f*** off you f***ing c***!

(b) ASKfm collection

Figure 1: Example content in the collections

tion approaches used in social media are complex, but in the end reduce to some combination of machine-assisted manual decisions (phase one) and automated decisions based on deploying a trained model (phase two). Operational decisions such as flagging and screening all posts from an account or massive reviewing of posts related to certain events [4, 6] are all results of applying previously trained models, which is also a form of deployment. Also, broadly applying the model to filter the content vastly reduces moderation burden when similar content is rapidly being published on the platform with the risk of falsely removal [4]. We claim no optimal for this specific simplified model in evaluating content moderation, but an initial effort for modeling the human-in-the-loop moderation process.

When applying the model to content moderation, however, we assume uniform review costs for all documents. This seems the best assumption given the short length of texts reviewed and what is known publicly about the cost structure of moderation [6].

In the next section, we describe our experimental setting for adapting and evaluating TAR for content moderation.

4. Experiment Design

Here we review the data sets, evaluation metric, and implementation details for our experiment.

4.1. Data Sets

We used two fully labeled and publicly available content moderation data sets with a focus on inappropriate user-generated content. The Wikipedia personal attack data set [32] consists of 115,737 Wikipedia discussion comments with labels obtained via crowdsourcing. An example of the comment is presented in Figure 1(a) Eight

annotators assigned one of five mutually exclusive labels to each document: Recipient Target, Third Party Target, Quotation Attack, Other Attack, and No Attack (our names). We defined three binary classification tasks corresponding to distinguishing Recipient Target, Third Party Target, or Other Attack from all other classes. (Quotation Attack had too low a prevalence.) A fourth binary classification task distinguished the union of all attacks from No Attack. A document was a positive example if 5 or more annotators put it in the positive class. Proportion of the positive class ranged from 13.44% to 0.18%.

The ASKfm cyberbullying dataset [29] contains 61,232 English utterance/response pairs, each of which we treated as a single document. An example of the conversation is presented in Figure 1(b). Linguists annotated both the poster and responder with zero or one of four mutually exclusive cyberbullying roles, as well as annotating the pair as a whole for any combination of 15 types of textual expressions related to cyberbullying. We treated these annotations as defining 23 binary classifications for a pair, with prevalence of the positive examples ranging from 4.63% to 0.04%.

For both data sets we refer to the binary classification tasks as *topics* and the units being classified as *documents*. Documents were tokenized by separating at punctuation and whitespace. Each distinct term became a feature. We used *log tf* weighting as the features for the underlying classification model. The value of a feature was 0 if not present, and else $1 + log(tf)$, where tf is the number of occurrences of that term in the document.

4.2. Algorithms and Workflow

Our experiments simulated a typical TAR workflow. The first training round is a seed set consisting of one random positive example (simulating manual input) and one random negative example. At the end of each round, a logistic regression model was trained and applied to the unlabeled documents. The training batch for the next round was then selected by one of three methods: a random sampling baseline, uncertainty sampling [52], or relevance feedback (top scoring documents) [51]. Variants of the latter two are widely used in eDiscovery [57]. Labels for the training batch were looked up, the batch was added to the training set, and a new model trained to repeat the cycle. Batches of size 100 and 200 were used and training continued for 80 and 40 iterations respectively, resulting in 8002 coded training documents at the end.

We implemented the TAR workflow in libact[5] [58], an open-source framework for active learning experiments. We fit logistic regression models using Vowpal Wabbit[6] with default parameter settings. Our experiment

framework is available on GitHub[7].

4.3. Evaluation

Our metric was total cost to reach 80% recall as described in Section 3. This was computed at the end of each training round as the sum of the number of training documents, plus the ideal second phase review cost as a penalty, which is the number of additional top-ranked documents (if any) needed to bring recall up to 80%. Ranking was based on sorting the non-training documents by probability of relevance using the most recent trained model. Note that we experimented with 80% recall as an example. However, the TAR workflow is capable of running with arbitrary recall target, such as 95% for systematic review [18, 56].

In actual TAR workflows, recall would be estimated from a labeled random sample. Since the cost of this sample would be constant across our experimental conditions we used an oracle for recall instead.

5. Results and Analysis

Our core finding was that, as in eDiscovery, active selection of which documents to review reduces costs over random selection. Figure 2 shows mean cost to reach 80% recall over 20 replications (different seed sets and random samples) for six representative categories. On all six categories, all TAR workflows within a few iterations beat the baseline of reviewing a random 80% of the data set (horizontal line labeled Manual Review).

The Wikipedia *Attack* category is typical of low to moderate prevalence categories ($p = 0.1344$). Uncertainty sampling strongly dominates both random sampling (too few positives chosen) and relevance feedback (too many redundant positives chosen for good training). Costs decrease uniformly with additional training. We plot 99% confidence intervals under the assumption that costs are normally distributed across replicates. Costs are not only higher for relevance feedback, but less predictable.

The ASKfm *Curse Exclusion* ($p = 0.0169$) and Wikipedia *Other attack* ($p = 0.0019$) category are typical low prevalence categories. Uncertainty sampling and relevance feedback act similarly in such circumstances: even top scoring documents are at best uncertainly positive. Average cost across replicates levels off and starts to increase after 44 iterations for uncertainty sampling and 45 iterations for relevance feedback. This is the point at which additional training no longer pays for itself by improving the ranking of documents. For this category (and typically) this occurs shortly before 80% recall is reached
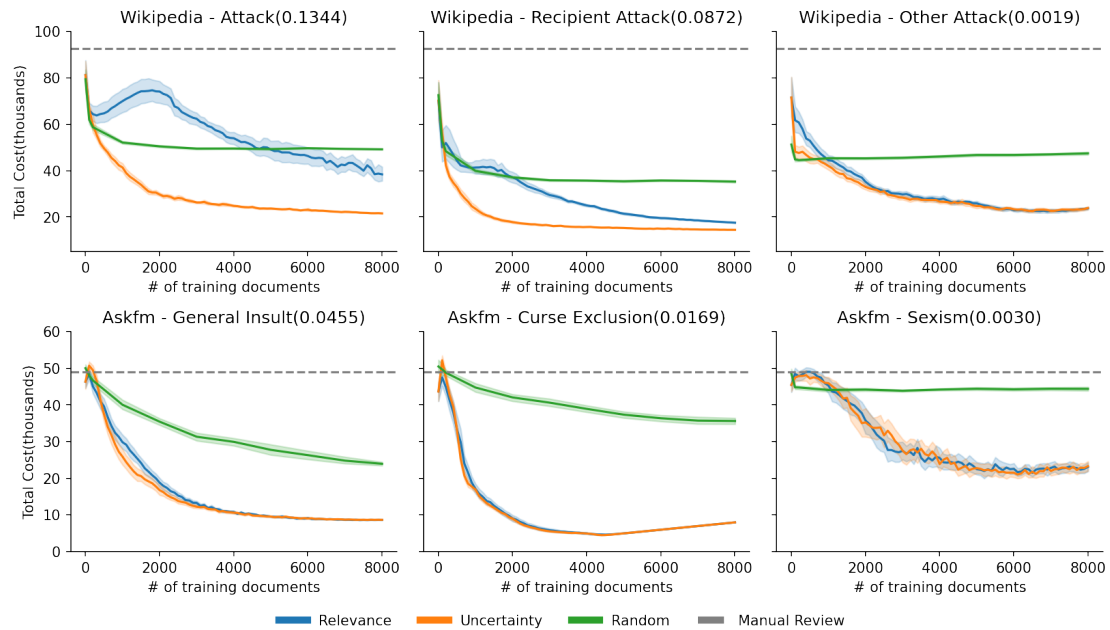
Figure 2: Total cost for TAR alternatives to identify 80% of positive documents for Wikipedia *Attack*, *Other Attack*, and *Recipient Attack*, and ASKfm *Curse Exclusion*, *General Insult*, and *Sexism* classifications. Values are averaged over 20 replicates, and a 99% confidence interval on costs is shown as shading around each curve. Horizontal line is cost to review a random 80% of the data set.

on the training data alone (iteration 48 for uncertainty sampling and iteration 52 for relevance feedback).

Task such as the ASKfm *Sexism* category ($p = 0.0030$) that deals with nuances in human languages requires more training data to produce a stable classifier. While obtaining training data by random sampling stops reducing the cost after the first iteration, uncertainty sampling and relevance feedback continue to take advantage of additional training data to minimize the cost and become more predictable.

Note that the general relationship between the prevalence of the task and the cost of reaching a certain recall target using TAR workflows is discussed Yang et al. [9].

Table 1 looks more broadly at the two datasets, averaging costs both over all topics and over 20 replicate runs for each topic for batch sizes of both 100 and 200 . By 20 iterations with batch size of 100 (2002 training documents), TAR workflows with both relevance feedback and uncertainty sampling significantly reduce costs versus TAR with random sampling. (Significance is based on paired t-tests assuming non-identical variances and making a Bonferroni correction for 72 tests.) All three TAR methods in turn dominate reviewing a random 80% of the dataset, which costs 92,590 for Wikipedia and 90,958 for ASKfm.

The improvement over cost plateaued after the train-

ing sets reached 5000 documents for ASKfm but continue for Wikipedia. Categories in Wikipedia ($p = 0.1344$ to 0.0018) are generally more frequent comparing to ASKfm ($p = 0.0463$ to 0.001), providing more advantage for training to identify more positive documents. Larger batch size slightly reduce the improvement as the underlying classifiers are retrained less frequently. In practice, the sizes are depending on the cost structure of reviewing and specific workflows in each organization. However, as the classifiers are frequently updated with more coded documents, the total cost would be reduced over the iterations.

Besides the overall cost reduction, Figure 3 shows a heatmap of mean precision across 20 replicates for batches 1 to 81 with batch size of 100, to give insight into the moderator experience of TAR workflows. Precision for relevance feedback starts high and declines very gradually. Uncertainty sampling maintains relatively constant precision. For the very low prevalence category Curse Exclusion we cut off the heatmap at 52 iterations for relevance feedback and 48 iterations for uncertainty sampling since on average 80% recall is obtained on training data alone by those iterations. For both categories, even applying uncertainty sampling that is intended to improve the quality of the classifier improves the batch precision over the random sampling be

Table 1
Total review cost to reach 80% recall. Values are averaged over all topics for a data set and 20 replicates. Percentages show relative cost reduction over the random sample training baseline. A * indicates that the difference is statistically significant over the random sample training baseline with 99% confidence by conducting paired t-test with Bonferroni correction.

| batch	# Train	ASKfm			Wikipedia		
		Random	Relevance	Uncertainty	Random	Relevance	Uncertainty
100	202	47685.53	*49833.73 (−4.50)	*50273.21 (−5.43)	52948.45	*60751.69 (−14.74)	52210.00 (1.39)
	1002	46327.93	*43329.31 (6.47)	*42723.12 (7.78)	49010.71	52931.28 (−8.00)	*39879.78 (18.63)
	2002	45139.15	*38179.79 (15.42)	*37938.19 (15.95)	47805.25	46673.34 (2.37)	*29387.06 (38.53)
	3002	44148.28	*34909.72 (20.93)	*34719.50 (21.36)	47065.66	*38964.91 (17.21)	*25676.82 (45.44)
	4002	43731.25	*33439.69 (23.53)	*32795.05 (25.01)	47234.75	*34408.14 (27.16)	*24202.29 (48.76)
	5002	43469.91	*32261.33 (25.78)	*31957.57 (26.48)	47125.79	*31267.88 (33.65)	*22746.94 (51.73)
	6002	42973.85	*31767.73 (26.08)	*31384.51 (26.97)	47300.02	*28945.59 (38.80)	*21922.42 (53.65)
	7002	42563.09	*30567.00 (28.18)	*30502.95 (28.33)	47086.42	*27356.89 (41.90)	*21301.92 (54.76)
	8002	42385.43	*30708.85 (27.55)	*30441.77 (28.18)	47106.34	*25949.51 (44.91)	*21144.28 (55.11)
200	202	47685.53	*49302.36 (−3.39)	*49339.93 (−3.47)	52948.45	*58866.41 (−11.18)	55747.35 (−5.29)
	1002	46327.93	45014.51 (2.84)	44733.10 (3.44)	49010.71	*55302.14 (−12.84)	*42896.71 (12.47)
	2002	45139.15	*40473.12 (10.34)	*39894.98 (11.62)	47805.25	49968.88 (−4.53)	*33981.56 (28.92)
	3002	44148.28	*37050.02 (16.08)	*36902.63 (16.41)	47065.66	42521.55 (9.65)	*28332.55 (39.80)
	4002	43731.25	*35310.13 (19.26)	*34888.22 (20.22)	47234.75	*37492.98 (20.62)	*25667.95 (45.66)
	5002	43469.91	*33690.33 (22.50)	*33519.15 (22.89)	47125.79	*34933.90 (25.87)	*24070.44 (48.92)
	6002	42973.85	*32425.25 (24.55)	*32612.13 (24.11)	47300.02	*33004.90 (30.22)	*22839.39 (51.71)
	7002	42563.09	*31488.77 (26.02)	*31813.08 (25.26)	47086.42	*31664.04 (32.75)	*22084.88 (53.10)
	8002	42385.43	*31198.75 (26.39)	*31171.80 (26.46)	47106.34	*29346.76 (37.70)	*21837.84 (53.64)
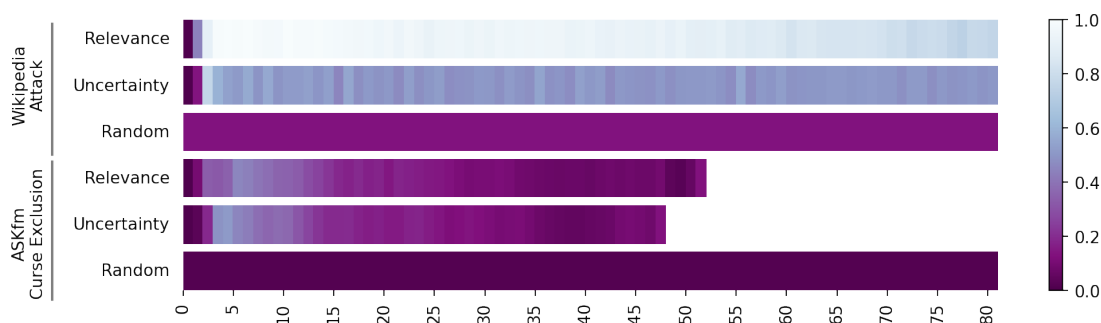
Figure 3: Precision in each batch for TAR workflows on Wikipedia *Attack*($p = 0.1344$) and ASKfm *Curse Exclusion*($p = 0.0169$) classifications. The x-axis shows the iteration number. A lighter color in an iteration block indicates higher precision.

a significant amount.

6. Summary and Future Work

Our results suggest that TAR workflows developed for legal review tasks may substantially reduce costs for content moderation tasks. Other legal workflow techniques, such as routing near duplicates and conversational threads in batches to the same reviewer, may be worth testing as well.

This preliminary experiment omitted complexities that should be explored in more detailed studies. Both content moderation and legal cases involve (at different time scales) streaming collection of data, and concomitant constraints on the time available to make a review decision. Batching and prioritization must reflect these constraints. Moderation in addition must deal with temporal variation in both textual content and the definitions of sensitive content, as well as scaling across many languages and cultures. As litigation and investigations become more international, these challenges may be faced in the law as well, providing opportunity for the legal and moderation fields to learn from each other.

References

[1] N. Duarte, E. Llanso, A. Loup, Mixed messages? the limits of automated social media content analysis, in: Conference on Fairness, Accountability and Transparency, PMLR, 2018, pp. 106–106.

[2] J. A. Gallo, C. Y. Cho, Social Media: Misinformation and Content Moderation Issues for Congress, Technical Report R46662, Congress Research Service, 2021. URL: https://crsreports.congress.gov/product/pdf/R/R46662.

[3] M. Ruckenstein, L. L. M. Turunen, Re-humanizing the platform: Content moderators and the logic of care, new media & society (2019) 1461444819875990.

[4] C. Consultants, Use of AI in Online Content Moderation, 2019. URL: https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.

[5] R. Gorwa, R. Binns, C. Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance, Big Data & Society 7 (2020) 2053951719897945.

[6] T. Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media, Yale University Press, 2018. URL: https://books.google.com/books?id=cOJgDwAAQBAJ.

[7] Fact.MR, Content Moderation Solutions Market Forecast, Trend Analysis & Competition Tracking - Global Market Insights 2019 to 2029, Technical Report FACT4522MR, Fact.MR, 2020. URL: https://www.factmr.com/report/4522/content-moderation-solutions-market.

[8] M. Surguy, International E-discovery: A Global Handbook of Law and Technology, Global Law and Business Limited, 2018. URL: https://books.google.com/books?id=pfK3swEACAAJ.

[9] E. Yang, D. D. Lewis, O. Frieder, On minimizing cost in legal document review workflows, in: Proceedings of the 21st ACM Symposium on Document Engineering), 2021.

[10] R. G. Maura, V. C. Gordon, Quantifying Success: Using Data Science to Measure the Accuracy of Technology-Assisted Review in Electronic Discovery, in: Data-Driven Law: Data Analytics and the New Legal Services, CRC Press, 2018, pp. 127–152.

[11] J. Baron, R. Losey, M. Berman, Perspectives on Predictive Coding: And Other Advanced Search Methods for the Legal Practitioner, American Bar Association, Section of Litigation, 2016. URL: https://books.google.com/books?id=TdJ2AQAACAAJ.

[12] J. Tredennick, TAR for Smart People, Catalyst Repository Systems, 2015.

[13] D. W. Oard, J. R. Baron, B. Hedin, D. D. Lewis, S. Tomlinson, Evaluation of information retrieval for e-discovery, Artificial Intelligence and Law 18 (2010) 347–386.

[14] A. Roegiest, G. V. Cormack, M. R. Grossman, C. Clarke, Trec 2015 total recall track overview, in: TREC, 2015.

[15] M. R. Grossman, G. V. Cormack, A. Roegiest, Trec 2016 total recall track overview., in: TREC, 2016.

[16] J. R. Baron, N. Payne, Dark archives and edemocracy: strategies for overcoming access barriers to the public record archives of the future, in: 2017 Conference for E-Democracy and Open Government (CeDEM), IEEE, 2017, pp. 3–11.

[17] I. J. Marshall, B. C. Wallace, Toward systematic review automation: a practical guide to using machine learning tools in research synthesis, Systematic reviews 8 (2019) 163.

[18] B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, C. H. Schmid, Semi-automated screening of biomedical citations for systematic reviews, BMC bioinformatics 11 (2010) 55.

[19] V. Bekkers, A. Edwards, D. de Kool, Social media monitoring: Responsive governance in the shadow of surveillance?, Government Information Quarterly 30 (2013) 335–342.

[20] A. Veglis, Moderation techniques for social media content, in: International Conference on Social Computing and Social Media, Springer, 2014, pp. 137–148.

[21] K. Langvardt, Regulating online content moderation, The Georgetown law journal 106 (2017) 1353.

[22] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, PloS one 14 (2019) e0221152.

[23] T. Xiang, S. MacAvaney, E. Yang, N. Goharian, ToxCCIn: Toxic content classification with interpretability, in: 11th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, 2021. URL: https://arxiv.org/abs/2103.01328.

[24] A. Holland, C. Bavitz, J. Hermes, A. Sellars, R. Budish, M. Lambert, N. Decoster, Intermediary liability in the united states, Network of Centers–Publixphere (2014).

[25] A. Halevy, C. C. Ferrer, H. Ma, U. Ozertem, P. Pantel, M. Saeidi, F. Silvestri, V. Stoyanov, Preserving integrity in online social networks, arXiv preprint arXiv:2009.10311 (2020).

[26] S. Akhtar, V. Basile, V. Patti, Modeling annotator perspective and polarized opinions to improve hate speech detection, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 8, 2020, pp. 151–154.

[27] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in:

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1668–1678.

[28] J. Pavlopoulos, P. Malakasiotis, I. Androutsopoulos, Deeper attention to abusive user content moderation, in: Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 1125–1135.

[29] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, V. Hoste, Automatic detection of cyberbullying in social media text, PloS one 13 (2018) e0203794.

[30] K. Reynolds, A. Kontostathis, L. Edwards, Using machine learning to detect cyberbullying, in: 2011 10th International Conference on Machine learning and applications and workshops, volume 2, IEEE, 2011, pp. 241–244.

[31] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International workshop on natural language processing for social media, 2017, pp. 1–10.

[32] E. Wulczyn, N. Thain, L. Dixon, Ex machina: Personal attacks seen at scale, in: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2017, pp. 1391–1399.

[33] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Eleventh international aaai conference on web and social media, 2017.

[34] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, N. Bhamidipati, Hate speech detection with comment embeddings, in: Proceedings of the 24th international conference on world wide web, ACM, 2015, pp. 29–30.

[35] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[36] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, International World Wide Web Conferences Steering Committee, 2016, pp. 145–153.

[37] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Predicting the type and target of offensive posts in social media, arXiv preprint arXiv:1902.09666 (2019).

[38] R. Kumar, A. N. Reganti, A. Bhatia, T. Maheshwari, Aggression-annotated corpus of hindi-english code-mixed data, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), 2018.

[39] G. K. Pitsilis, H. Ramampiaro, H. Langseth, Detecting offensive language in tweets using deep learning, arXiv preprint arXiv:1801.04433 (2018).

[40] S. Sotudeh, T. Xiang, H.-R. Yao, S. MacAvaney, E. Yang, N. Goharian, O. Frieder, Guir at semeval-2020 task 12: Domain-tuned contextualized models for offensive language detection, arXiv preprint arXiv:2007.14477 (2020).

[41] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), arXiv preprint arXiv:1903.08983 (2019).

[42] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020), arXiv preprint arXiv:2006.07235 (2020).

[43] R. Binns, M. Veale, M. Van Kleek, N. Shadbolt, Like trainer, like bot? inheritance of bias in algorithmic content moderation, in: International Conference on Social Informatics, Springer, 2017, pp. 405–415.

[44] L. Dixon, J. Li, J. Sorensen, N. Thain, L. Vasserman, Measuring and mitigating unintended bias in text classification, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 67–73.

[45] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, arXiv preprint arXiv:1908.09635 (2019).

[46] D. Link, B. Hellingrath, J. Ling, A human-is-the-loop approach for semi-automated content moderation., in: ISCRAM, 2016.

[47] N. M. Pace, L. Zakaras, Where the money goes: Understanding litigant expenditures for producing electronic discovery, RAND Corporation, 2012.

[48] M. Bagdouri, W. Webber, D. D. Lewis, D. W. Oard, Towards minimizing the annotation cost of certified text classification, in: CIKM 2013, ACM, 2013, pp. 989–998.

[49] G. V. Cormack, M. R. Grossman, Autonomy and reliability of continuous active learning for technology-assisted review, arXiv preprint arXiv:1504.06868 (2015).

[50] B. Settles, Active learning literature survey (2009).

[51] J. Rocchio, Relevance feedback in information retrieval, The Smart retrieval system-experiments in automatic document processing (1971) 313–323.

[52] D. D. Lewis, W. A. Gale, A sequential algorithm for training text classifiers, in: SIGIR 1994, 1994, pp. 3–12.

[53] G. V. Cormack, M. R. Grossman, Engineering Quality and Reliability in Technology-Assisted Review,

in: SIGIR, ACM Press, Pisa, Italy, 2016, pp. 75–84. URL: http://dl.acm.org/citation.cfm?doid=2911 451.2911510. doi:10.1145/2911451.2911510, 00024.

[54] D. D. Lewis, E. Yang, O. Frieder, Certifying one-phase technology-assisted reviews (2021).

[55] E. Yang, D. D. Lewis, O. Frieder, Heuristic stopping rules for technology-assisted review, in: Proceedings of the 21st ACM Symposium on Document Engineering, 2021.

[56] D. Li, E. Kanoulas, When to stop reviewing in technology-assisted reviews: Sampling from an adaptive distribution to estimate residual relevant documents, ACM Transactions on Information Systems (TOIS) 38 (2020) 1–36.

[57] G. F. Cormack, M. F. Grossman, Evaluation of machine-learning protocols for technology-assisted review in electronic discovery, SIGIR 2014 (2014) 153–162. doi:10.1145/2600428.2609601.

[58] Y.-Y. Yang, S.-C. Lee, Y.-A. Chung, T.-E. Wu, S.-A. Chen, H.-T. Lin, libact: Pool-based Active Learning in Python, Technical Report, National Taiwan University, 2017. URL: https://github.com/ntu cllab/libact, available as arXiv preprint https://arxiv.org/abs/1710.00379.