A human-ML collaboration framework for improving video content reviews

Meghana Deodhar, Xiao Ma, Yixin Cai, Alex Koes, Alex Beutel and Jilin Chen

Google, USA

Abstract

We deal with the problem of localized in-video taxonomic human annotation in the video content moderation domain, where the goal is to identify video segments that violate granular policies, e.g., community guidelines on an online video platform. High quality human labeling is critical for enforcement in content moderation. This is challenging due to the problem of information overload - raters need to apply a large taxonomy of granular policy violations with ambiguous definitions, within a limited review duration to relatively long videos. Our key contribution is a novel human-machine learning (ML) collaboration framework aimed at maximizing the quality and efficiency of human decisions in this setting - human labels are used to train segment-level models, the predictions of which are displayed as "hints" to human raters, indicating probable regions of the video with specific policy violations. The human verified/corrected segment labels can help refine the model further, hence creating a human-ML positive feedback loop. Experiments show improved human video moderation decision quality, and efficiency through more granular annotations submitted within a similar review duration, which enable a 5-8% AUC improvement in the hint generation models.

Keywords

human computation, machine learning, video content moderation, ranking

1. Introduction

The importance of content moderation on online video platforms such as TikTok, YouTube or Instagram is growing [1, 2]. These platforms strive to accurately detect the presence of policy violations within the video, which drive enforcement actions, e.g., the video can be taken down. Given the complexity of this problem, content moderation relies heavily on human judgement and employs large teams of content moderators to perform reviews. Since human annotations directly lead to high stakes decisions, such as content take downs, the quality of the annotations is critical.

For content moderation decisions there is a growing need for transparency in detected policy violations to provide feedback to content creators [3]. This motivates a in-video taxonomic annotation task, where the goal is to provide *localized* and *fine-grained policy-specific* annotations, i.e., both the time regions (video segments) and the exact policies violated, which inform downstream video content moderation decisions. To cover the spectrum of potential violations, policy playbooks typically contain hundreds of fine-grained policies. This large space of policy violations can be organized as a taxonomy of broad categories such as Profanity, Violence, Nudity, etc., each of which contains several granular violations. For instance, Violence could include a range of granular classes such as animal abuse or graphic violence in video games. The class definitions are complex, ambiguous and often require nuanced judgment to apply, e.g., graphic violence. New policy classes may be added over time as well, e.g., Covid anti-vaccination. Moreover, there is a class imbalance issue - some egregious violations may be very rare.

Our goal is to maximize the quality and efficiency of the complex, granular, localized policy annotations task, hence leading to the correct video level enforcement decision. We achieve this by tackling the key issue of "information overload" faced by raters in providing high quality annotations, where 1) the sheer volume of videos on large online video platforms means raters only have limited review time per video; and 2) the large taxonomy of policies makes it hard for raters to recall the complete set of all granular violations for every video region they watch in the limited review duration.

We propose a human-ML collaboration framework to maximize human ratings quality and efficiency by addressing "information overload". We train models on granular rater annotations to predict policy violations, which are then combined with innovative front-end elements in the rating tool to provide "hints" to assist raters. We borrow from information retrieval literature and use ranking mechanisms for identifying the most useful and succinct set of hints. In experiments, we show that this enables raters to efficiently label policy violations more correctly and comprehensively. The human interactions

with model hints pave the way for leveraging human feedback to improve the underlying ML models.

2. Related Work

The crowd sourcing literature is very rich in the application of human annotations to perform a variety of tasks such as text processing [4, 5, 6], audio transcription [7], taxonomy creation [8], and social media analysis [9, 10]. Although there is existing literature on video annotation, it is primarily focused on identifying actions or labeling entities easily distinguishable by humans using visual information only, e.g., high jump, thunderstorms. The primary goal of these tasks is to create large datasets for facilitating Machine Learning/Perception applications [11, 12], e.g., the YouTube-8m dataset [13]. This is very different from our set up, where raters annotate granular, ambiguously defined policies using multi-modal signals - video, audio and text, from the transcript, and the video title and description. The recent emergence of crowd sourcing literature on content moderation primarily covers textual content such as user comments [14, 15], however there has been little focus on video moderation tasks.

Human-ML collaboration is an emerging area of research with two main categories of work:

ML-Assisted Human Labeling: ML-assistance through predictions and explanations has been used to improve the quality of human decisions in several domains [16, 17, 18], including content moderation [15, 14]. Crossmod [15], for instance, uses a model trained on historic cross-community moderation decisions to enable Reddit human moderators to find more violations. Interactive ML-assistance, which we leverage in Section 3.1.2, is used by Bartolo et al. [19] to assist human annotators to develop adversarial examples for improving a natural language question answering model. ML-assistance has been shown to also improve the efficiency of the human labeling task [20, 15, 21]. Our work aims to exploit both the human labeling quality and efficiency benefits.

Improving ML-models Through Human Annotations: Human annotations are useful in constructing hybrid human-ML systems that leverage the complementary strengths of both to improve the performance of ML models [22, 23]. Existing work on active learning [24, 25] shows that strategically sampling data points can reduce human workload, but the purpose is to improve machine learning models instead of assisting raters. Recent work on explainable active learning (XAL) [26] has called for better designing for the human experience in the human-AI interface.

Our work shows that it is possible to achieve model improvements and assist human raters, bridging the gap between ML-assisted human labeling and active learn-
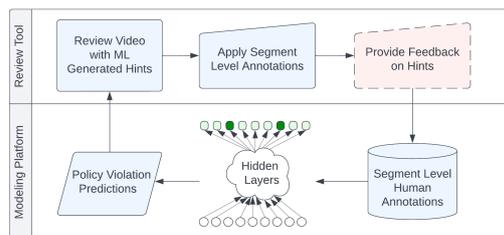

Figure 1: Human-ML collaboration set up.

ing. The novelty of our approach is that the models are re-trained continuously on the output of the human annotation task, which they provide assistance for, constructing a positive feedback loop between humans and models. In this collaborative framework, we have the opportunity to improve both modeling and human rater performance.

Information overload, which we encounter in our content moderation setting, is a well studied problem that reduces the effectiveness of a human's decision making ability [27, 28]. To address this, we build on the intuition that humans find it easier to verify or correct suggestions rather than produce new annotations from scratch [29, 5]. Our ML-assistance proposal strives to select the most informative but succinct ML-based "hints" to surface to raters by drawing on the information retrieval and ranking literature. ML-based ranking has been shown to reduce information overload effectively in electronic messaging [30] and social media [31, 32]. We draw inspiration from the learning to rank idea [33] to reduce information overload for raters.

3. Proposed Human-ML Collaboration Framework

The main contribution of this paper is the human-ML collaboration framework visualized in Figure 1. We use the predictions of ML models to provide assistance to human raters and evaluate the effectiveness of different user interfaces for the ML-assistance. Since the policy violation prediction task is hard for ML models, the feedback from human raters is useful to improve the models. ML-assistance enables raters to provide segment level annotations more efficiently leading to more ground truth to train/update the ML models. Additionally, we can enable raters to interact with the ML hints (accept/reject), providing direct feedback to refine the model, establishing a positive human-ML feedback loop.

3.1. ML-Assisted Human Reviews

As discussed in the introduction, raters face an information overload problem due to a combination of granular, complex policy definitions and limited time per video. Often, raters need to use their judgement to decide what parts of the video to watch to identify violating segments, resulting in fleeting violations being missed, or inconsistencies across raters watching different sections of the same video.

It is intuitive that raters would benefit from pointers to likely unsafe regions within a video, labeled with the exact policies being violated. Even if not completely precise, this will enable them to optimize their review bandwidth by focusing on potentially more relevant regions, making them less likely to miss violations. We achieve this by training per policy ML models and transforming their predictions into "hints" provided to raters, described in more detail in Section 3.1.2. We tune the models to be high recall to minimize uncaught violations, while relying on human judgement to improve the precision of the labeled violations.

3.1.1. Segment Level Model Training

To train segment level policy violation models we frame the following modeling problem - given multi-modal features for a fixed length video segment, predict whether the segment contains specific policy violations. We generate training datasets by extracting per-frame visual and audio features from the human labeled violation region. The visual and audio features are the dense embedding layers of standard convolutional network image semantic similarity [34] and audio classification models [35] respectively. Based on empirical evidence, we select flat concatenation to aggregate the frame features over a segment, versus average/max pooling. The final model we train is a multi-label DNN model with the aggregated frame-level visual and audio features as input, where each label corresponds to a fine grained policy violation. We use MultiModal Versatile Networks [36] during model training to learn a better representation for audio and visual features for our classification task, which further improves model performance. We use a sliding window approach to utilize the trained model to generate prediction scores per policy violation for a fixed length segment starting at each frame of the video. Using a window of frame size n and stride of 1 frame, we produce model scores for segments with start and end frames [0 to n-1], [1 to n], and so on until the end of the video, padding with empty features to fit the segment length for the last n frames.

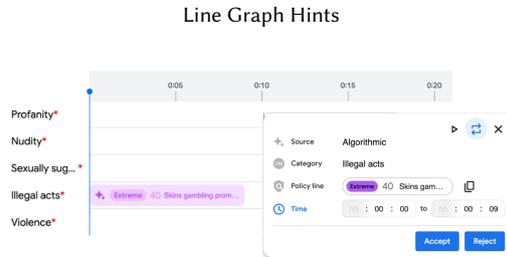**Figure 2:** V1 Continuous Line Graph for Specific Policy Hints

Line Graph Hints

Figure 3: V2 Pre-Populated Segments from ML Models

Pre-Populated Segments

3.1.2. Techniques to Provide ML-Assistance

We proceed to develop ways to use model predictions to most effectively assist human reviews, and provide details on two different designs (V1 and V2).

V1 Hints: Continuous Line Graphs. For video annotations, it is standard to display the video itself with playback controls and additional information in the form of a timeline [37]. For V1, we display the ML predictions as a line graph across the entire timeline of the video. The user interface is demonstrated in Figure 2. Raters can examine the line graphs and jump to the point in the video where a peak (policy violation) occurs.

While we have model predictions for hundreds of granular policies, due to visual clutter, we only display plots for a small subset of the most frequent policies. Raters also don't have the ability to provide feedback to improve model predictions.

V2 Hints: Towards a Scalable and Interactive-ML Assistance UI. In V2, we borrow elements from recommender systems [38, 39] to develop a more *scalable* and *interactive* interface (see Figure 3), where we pre-populate video segments that may contain policy violations detected by machine learning models.

To generate video segments, as shown in Figure 3, we introduce an algorithm to binarize continuous model scores per policy into discrete segments and use a ranking algorithm to recommend the most useful segments to raters. For simplicity, we used a threshold-based algorithm. The threshold selection constitutes a tradeoff between the precision and recall, where precision cap-

tures the utility of the predicted segments to raters and recall captures the comprehensive coverage of all video violations. The algorithm chooses a threshold maximizing recall, while maintaining a minimum precision (40% based on user studies). The regions of the video where the ML scores are above this chosen threshold are displayed as predicted policy violating segments to raters. Several heuristics are then applied to maximize segment quality, e.g., we merge segments that are close to avoid visual clutter (<3% of the whole video length apart).

Finally, to reduce information overload, we borrow from the learning to rank concept [33] to rank candidate segments and limit the number of displayed segments. The ranking algorithm prioritizes segments based on the max score across segment frames, and egregiousness of the predicted policies. We then select the top N segments to display to raters, with N selected through user studies. We pre-populate each ML suggested segment in the video timeline UI as seen in Figure 3. Raters can choose to accept or reject the suggested segments. These logged interactions can be used to provide feedback to improve the ML models.

3.2. Human Feedback to ML-Models

The ambiguous and fluid policy definitions along with a changing distribution of videos on online platforms poses a challenge for building robust models to accurately predict policy violations for providing ML-assistance. We hence continuously need more data and human feedback to improve the models. We show that ML-assistance increases the total number of segment labels submitted vs. no assistance. The labels in turn can serve as new ground truth for continuously re-training the models and improving performance. Additionally, with the reject button shown in V2 we collect clean negative labels; earlier, we had only "weak" negatives from videos where no violations were annotated [1]. Further exploiting the human feedback in combination with active learning strategies is an area of future work.

4. Experiments

Our proposed methodology is evaluated using raters from the author's organization that regularly perform video content moderation reviews on live queues. Raters are separated into two pools: experts and generalists, with 150 and 400 raters respectively. The expert pool is a group of more experienced quality assurance (QA) raters with demonstrably better decision quality over a long period of time. Since their focus is QA, they don't have fixed

---

[1]Even if no violations were annotated, they could still be present in video segments the rater did not watch, hence the negative labels inferred are weak/noisy.

Treatment	Precision	Recall	# Videos
V1 vs. Baseline	+9.82%	+1.37%	3456
V1 + V2 vs. V1	+9.97%	+5.64%	2914

Table 1
Relative impact on live traffic (Dataset 1)

productivity targets as generalists do. They can hence spend more time reviewing each video comprehensively, leading to higher decision quality. In our experiment setup, each video in an evaluation dataset is independently reviewed by 1 expert and 2 generalist raters. We use the labels from expert raters on the dataset as the ground truth to evaluate the 2 sets of generalist rater decisions.

4.1. Experimental Setup

4.1.1. Datasets

Our two evaluation datasets contain videos viewed on a large online video platform: (1) "Live traffic" dataset: Sampled from live traffic, hence containing a very low proportion of policy violating videos; (2) "Affected slice" dataset, sampled from live traffic and filtered to only videos with ML-hints present, containing 13-20% policy violating videos.

4.1.2. Metrics

Our human ratings quality metric is calculated at the video-level and conveys the correctness of the final, binary content moderation decision, e.g., take down or not. We compute the precision (P), recall (R), and disagreement rate for each of the 2 sets of generalist's video-level decisions. We consider the expert decision as ground truth and report the averaged values across the 2 sets. On live traffic datasets, we use P/R over standard inter-rater disagreement metrics due to the high class imbalance.

For rater efficiency, we measure: (1) percentage of policy violating videos where raters provide segment annotations. (2) number of segment annotations submitted by raters per video (3) average review duration per video.

4.2. Results

4.2.1. Rater Quality Improvements

We conduct experiments on both "live traffic" and "affected slice" datasets, with the baseline as a review process without ML-hints. Tables 1 and 2 compare the ratings quality metrics of our proposed V1 (line plot of model scores) ML-assistance treatment relative to the baseline, and evaluate the incremental benefit of the V2 (pre-populated segments) treatment over V1, in the V1 + V2 vs. V1 row.

Treatment	Precision	Recall	Disagreement%	# Videos
V1 vs. Baseline	+4.24%	+3.64%	-15.71%	3319
V1 + V2 vs. V1	+7.02%	+14.30%	-32.27%	682

Table 2
Relative impact on affected slice (Dataset 2)

From the live traffic results in Table 1, we see that V1 shows an improvement in precision and recall over the baseline, driven by the improvement on the affected slice as seen in Table 2.

We also see large rater quality gains of V2 over V1 on both live traffic and affected slice datasets. The segmentation and ranking algorithms in V2 allows us to overcome the scalability limitation of V1 and expand the number of granular policies covered by model hints from 7 to 18. Specifically for violence related violations, we see a 35% relative recall gain over V1 by expanding policies with ML-hints from 2 to 9. The V2 design can be scaled to cover hundreds of policies in future versions by dynamically surfacing the most relevant violating segments, further improving recall.

4.2.2. Rater Efficiency Improvements

We observe reduced review duration on policy violating videos with V1 hints vs. without, with more efficiency benefits on longer videos as expected; -14% on videos longer than 10 minutes and -20% on videos longer than 30 minutes. With V2, relative to V1, we see a 3% increase in review duration, but it is traded off by a 9% increase in the percentage of policy violating videos with exact segments annotated, and a 24% increase in the number of segment annotations submitted per video.

4.2.3. Interactive ML-Assistance Metrics

Isolating the precision of the ML-assisted segments, we see raters accepting 35% of ML generated hints, which is in line with the 40% precision constraint we chose when converting model scores into discrete segments.

4.2.4. Model Quality Improvements

Since the introduction of V1 hints, we see significant model performance improvements with more human labels collected within a 3 month period on specific policy areas.

5. Discussion

One of the potential risks of our proposed human-ML collaboration framework is automation bias [40], where a rater's over-reliance on ML-assistance can result in (i) blind-spots due to humans missing violations and (ii)

Policy Area	AUCPR	# Positive Labels
Sexually Suggestive	+5.9%	+12.7%
Nudity	+4.9%	+12.7%
Illegal Acts	+8.6%	+12.1%

Table 3
Model Quality Improvements

raters accepting model hints without verification. Our video-level ratings quality evaluation metrics are robust to this since the ground truth comes from expert (QA) raters who review videos comprehensively, looking beyond ML-hints. In practice, we observe little evidence of both (i) and (ii). 56% of the violation segments submitted by raters in the V2 setup are organically created, i.e., don't overlap with pre-populated hint segments. The segment acceptance rate is 35%, aligned with our segmentation model precision tuning point of 40%, indicating that raters are verifying and rejecting false positive hints at the expected rate. We could mitigate the risk of (ii) further by enforcing that at least some percentage of hint segments/video is actually watched or by surfacing the model's confidence in the predicted hint to raters. To ensure robust evaluation of model quality, the AUC improvements in Section 4.2.4 are evaluated on a set of labels collected without model generated segments.

6. Future work

For content moderation to scale to the size of online platforms, it is necessary to take model-based enforcement action. We would like to explore the relation between improved ground truth and improvement of automated, model based enforcement. Leveraging active learning strategies in combination with utilizing rater feedback on model generated segments to show further quality improvements in the models is another open area of research. Finally, we will explore multi-armed bandits to balance active learning based *exploration* for model improvement with model *exploitation* for providing high quality ML-assistance [41].

This paper used content moderation as the test bed for our human-ML collaboration proposal. However, it is a more generalized framework that applies to the problem of granular, localized video annotation encountered in various other industry applications such as identifying products/brands in videos to inform the placement of

relevant ads, which we would like to explore further.

References

[1] K. Gogarty, Hate speech and misinformation proliferate on meta products, with 13,500 policy violations documented in the past year alone, https://www.mediamatters.org/facebook/hate-speech-and-misinformation-proliferate-meta-products-13500-policy-violations, 2022.

[2] S. Mellor, Tiktok slammed for videos sharing false information about russia's war on ukraine, https://fortune.com/2022/03/21/tiktok-misinformation-ukraine/, 2022.

[3] S. Jhaver, A. Bruckman, E. Gilbert, Does transparency in moderation really matter? user behavior after content removal explanations on reddit, Proc. ACM Hum.-Comput. Interact. 3 (2019). URL: https://doi.org/10.1145/3359252. doi:10.1145/3359252.

[4] A. Kittur, E. H. Chi, B. Suh, Crowdsourcing user studies with mechanical turk, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 453–456. URL: https://doi.org/10.1145/1357054.1357127. doi:10.1145/1357054.1357127.

[5] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, K. Panovich, Soylent: A word processor with a crowd inside, in: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, UIST '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 313–322. URL: https://doi.org/10.1145/1866029.1866078. doi:10.1145/1866029.1866078.

[6] C. Hu, B. B. Bederson, P. Resnik, Y. Kronrod, Monotrans2: A new human computation system to support monolingual translation, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 1133–1136. URL: https://doi.org/10.1145/1978942.1979111. doi:10.1145/1978942.1979111.

[7] W. Lasecki, C. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, J. Bigham, Real-Time Captioning by Groups of Non-Experts, Association for Computing Machinery, New York, NY, USA, 2012, p. 23–34. URL: https://doi.org/10.1145/2380116.2380122.

[8] L. B. Chilton, G. Little, D. Edge, D. S. Weld, J. A. Landay, Cascade: Crowdsourcing taxonomy creation, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013, pp. 1999–2008.

[9] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, P. Tolmie, Crowdsourcing the annotation of rumourous conversations in social media, in: Proceedings of the 24th international conference on World Wide Web, 2015, pp. 347–353.

[10] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of twitter abusive behavior, in: Twelfth International AAAI Conference on Web and Social Media, 2018.

[11] H. Zhao, A. Torralba, L. Torresani, Z. Yan, Hacs: Human action clips and segments dataset for recognition and temporal localization, 2017. URL: https://arxiv.org/abs/1712.09374. doi:10.48550/ARXIV.1712.09374.

[12] M. Z. Trujillo, M. Gruppi, C. Buntain, B. D. Horne, The mela bitchute dataset, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 16, 2022, pp. 1342–1351.

[13] S. Abu-El-Haija, N. Kothari, J. Lee, A. P. Natsev, G. Toderici, B. Varadarajan, S. Vijayanarasimhan, Youtube-8m: A large-scale video classification benchmark, in: arXiv:1609.08675, 2016. URL: https://arxiv.org/pdf/1609.08675v1.pdf.

[14] V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, C. Tan, Human-ai collaboration via conditional delegation: A case study of content moderation, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, Association for Computing Machinery, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3491102.3501999. doi:10.1145/3491102.3501999.

[15] E. Chandrasekharan, C. Gandhi, M. W. Mustelier, E. Gilbert, Crossmod: A cross-community learning-based system to assist reddit moderators, Proc. ACM Hum.-Comput. Interact. 3 (2019). URL: https://doi.org/10.1145/3359276. doi:10.1145/3359276.

[16] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, L. M. Vardoulakis, A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy, Association for Computing Machinery, New York, NY, USA, 2020, p. 1–12. URL: https://doi.org/10.1145/3313831.3376718.

[17] V. Lai, C. Tan, On human predictions with explanations and predictions of machine learning models: A case study on deception detection, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery,

New York, NY, USA, 2019, p. 29–38. URL: https://doi.org/10.1145/3287560.3287590. doi:10.1145/3287560.3287590.

[18] J. Park, R. Krishna, P. Khadpe, L. Fei-Fei, M. Bernstein, Ai-based request augmentation to increase crowdsourcing participation, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 7 (2019) 115–124. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/5282.

[19] M. Bartolo, T. Thrush, S. Riedel, P. Stenetorp, R. Jia, D. Kiela, Models in the loop: Aiding crowdworkers with generative annotation assistants, CoRR abs/2112.09062 (2021). URL: https://arxiv.org/abs/2112.09062. arXiv:2112.09062.

[20] Z. Ashktorab, M. Desmond, J. Andres, M. Muller, N. N. Joshi, M. Brachman, A. Sharma, K. Brimijoin, Q. Pan, C. T. Wolf, E. Duesterwald, C. Dugan, W. Geyer, D. Reimer, Ai-assisted human labeling: Batching for efficiency without overreliance, Proc. ACM Hum.-Comput. Interact. 5 (2021). URL: https://doi.org/10.1145/3449163. doi:10.1145/3449163.

[21] S. Anjum, A. Verma, B. Dang, D. Gurari, Exploring the use of deep learning with crowdsourcing to annotate images, Human Computation 8 (2021) 76–106. doi:10.15346/hc.v8i2.121.

[22] I. Arous, J. Yang, M. Khayati, P. Cudré-Mauroux, OpenCrowd: A Human-AI Collaborative Approach for Finding Social Influencers via Open-Ended Answers Aggregation, Association for Computing Machinery, New York, NY, USA, 2020, p. 1851–1862. URL: https://doi.org/10.1145/3366423.3380254.

[23] J. W. Vaughan, Making better use of the crowd: How crowdsourcing can advance machine learning research, Journal of Machine Learning Research 18 (2018) 1–46. URL: http://jmlr.org/papers/v18/17-234.html.

[24] B. Settles, Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[25] Y. Yang, Z. Ma, F. Nie, X. Chang, A. G. Hauptmann, Multi-class active learning by uncertainty sampling with diversity maximization, International Journal of Computer Vision 113 (2015) 113–127.

[26] B. Ghai, Q. V. Liao, Y. Zhang, R. Bellamy, K. Mueller, Explainable active learning (xal): An empirical study of how local explanations impact annotator experience, 2020. URL: https://arxiv.org/abs/2001.09219. doi:10.48550/ARXIV.2001.09219.

[27] M. J. Eppler, J. Mengis, The concept of information overload-a review of literature from organization science, accounting, marketing, mis, and related disciplines (2004), Kommunikationsmanagement im Wandel (2008) 271–305.

[28] Y.-C. Chen, R.-A. Shang, C.-Y. Kao, The effects of information overload on consumers' subjective state towards buying decision in the internet shopping environment, Electronic Commerce Research and Applications 8 (2009) 48–58.

[29] H. Hu, L. Xie, Z. Du, R. Hong, Q. Tian, One-bit supervision for image classification (2020). URL: https://arxiv.org/abs/2009.06168. doi:10.48550/ARXIV.2009.06168.

[30] R. M. Losee Jr, Minimizing information overload: the ranking of electronic messages, Journal of Information Science 15 (1989) 179–189.

[31] K. Koroleva, A. J. Bolufé Röhler, Reducing information overload: Design and evaluation of filtering & ranking algorithms for social networking sites (2012).

[32] J. Chen, R. Nairn, L. Nelson, M. Bernstein, E. Chi, Short and tweet: experiments on recommending content from information streams, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2010, pp. 1185–1194.

[33] A. Karatzoglou, L. Baltrunas, Y. Shi, Learning to rank for recommender systems, in: Proceedings of the 7th ACM Conference on Recommender Systems, 2013, pp. 493–494.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, 2015. URL: https://arxiv.org/abs/1512.00567. doi:10.48550/ARXIV.1512.00567.

[35] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, K. W. Wilson, CNN architectures for large-scale audio classification, CoRR abs/1609.09430 (2016). URL: http://arxiv.org/abs/1609.09430. arXiv:1609.09430.

[36] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, A. Zisserman, Self-supervised multimodal versatile networks, 2020. URL: https://arxiv.org/abs/2006.16228. doi:10.48550/ARXIV.2006.16228.

[37] C. Vondrick, D. Ramanan, D. Patterson, Efficiently scaling up video annotation with crowdsourced marketplaces, in: European Conference on Computer Vision, Springer, 2010, pp. 610–623.

[38] I. Portugal, P. Alencar, D. Cowan, The use of machine learning algorithms in recommender systems: A systematic review, Expert Systems with Applications 97 (2018) 205–227.

[39] K. Swearingen, R. Sinha, Interaction design for recommender systems, in: Designing Interactive Systems, volume 6, Citeseer, 2002, pp. 312–334.

[40] L. J. Skitka, K. L. Mosier, M. Burdick, Does automation bias decision-making?, International Journal of Human-Computer Studies 51 (1999) 991–1006.

[41] J. McInerney, B. Lacker, S. Hansen, K. Higley, H. Bouchard, A. Gruson, R. Mehrotra, Explore, exploit, and explain: personalizing explainable recommendations with bandits, in: Proceedings of the 12th ACM conference on recommender systems, 2018, pp. 31–39.