3D4ALL: Toward an Inclusive Pipeline to Classify 3D Contents

Nahyun Kwon[a], Chen Liang[a] and Jeeeun Kim[a]

[a]HCIED Lab, Texas A&M University

Abstract

Algorithmic content moderation manages an explosive number of user-created content shared online everyday. Despite a massive number of 3D designs that are free to be downloaded, shared, and 3D printed by the users, detecting sensitivity with transparency and fairness has been controversial. Although sensitive 3D content might have a greater impact than other media due to its possible reproducibility and replicability without restriction, prevailed unawareness resulted in proliferation of sensitive 3D models online and a lack of discussion on transparent and fair 3D content moderation. As the 3D content exists as a document on the web mainly consisting of text and images, we first study the existing algorithmic efforts based on text and images and the prior endeavors to encompass transparency and fairness in moderation, which can also be useful in a 3D printing domain. At the same time, we identify 3D specific features that should be addressed to advance a 3D specialized algorithmic moderation. As a potential solution, we suggest a human-in-the-loop pipeline using augmented learning, powered by various stakeholders with different backgrounds and perspectives in understanding the content. Our pipeline aims to minimize personal biases by enabling diverse stakeholders to be vocal in reflecting various factors to interpret the content. We add our initial proposal for redesigning metadata of open 3D repositories, to invoke users' responsible actions of being granted consent from the subject upon sharing contents for free in the public spaces.

Keywords

3D printing, sensitive contents, content moderation

1. Introduction

To date, many social media platforms observed an explosive number of user-created content posted everyday from Twitter to YouTube to Instagram and more. Following the acceleration of online contents which becomes even faster partly due to COVID-19, it has also become easier for people to access sensitive content that may not be appropriate for the general purpose. Owing to the scale of these content and users' abilities to share and repost them in a flash, it becomes extremely costly to detect the sensitive content solely by manual work. Current social media platforms have adopted various (semi)automated content moderation methods including a deep learning-based classification (e.g., Microsoft Azure Content Moderator [1], DeepAI's Nudity Detection API [2], Amazon Rekognition Content Moderation [3]).

Meanwhile, since desktop 3D printers have been flooded into the consumer market, 3D printing specific social platforms such as Thingiverse [4] have also gained popularity, contributing to the proliferation of shared 3D

Joint Proceedings of the ACM IUI 2021 Workshops, April 13-17, 2021, College Station, USA

nahyunkwon@tamu.edu (N. Kwon); cltamu@tamu.edu (C. Liang); jeeeun.kim@tamu.edu (J. Kim)

https://nahyunkwon.github.io/ (N. Kwon); http://www.jeeeunkim.com/ (J. Kim)

0000-0002-2332-0352 (N. Kwon); 0000-0003-1645-2397 (C. Liang); 0000-0002-8915-481X (J. Kim)

contents that are easily downloadable and replicable among community users. Despite a massive number of 3D contents shared for free to date—As of 2020 2Q, there are near 1.8 million 3D models available for download, excluding empty entries due to post deletion—, there has been relatively little attention to sensitive 3D contents. This might result in not only a lack of a dataset to be used as a bench mark, but also a lack of discussion on fair rationales to be utilized in building a algorithmic 3D content moderation that integrates everyone's perspectives with a different background. Along with significant advances in technology of machine mechanisms and materials (e.g., 3D printing in metals), the 3D printing community may present an even greater impact from the spread of content due to its limitless potential for replication and reproduction. In view of various stakeholders who have different perspectives in consuming and interpreting contents—from K-12 teachers who may seek 3D files online to design curricula to artists who depict their creativity in digitized 3D sculptures—, moderating 3D content with fairness becomes more challenging. 3D contents online often consist of images and text that are possibly useful to adopt existing moderation schemes including text (e.g., [5, 6, 7, 8]) or image based (e.g., [9, 10, 11]) approaches. However, there exist 3D printing specific features (e.g., print support to avoid overhangs, uni-colored outcome, segmented in parts, etc.) that may prevent direct adoption of those schemes, requiring further consideration about implementing advanced 3D content moderation techniques.

In this work, we first study the existing content moderation efforts that has potential to be used in 3D content moderationand discuss shared concerns in examining transparency and fairness issues in algorithmic content moderation. As a potential solution, we propose a semi-automated human-in-the-loop validation pipeline using augmented learning that incrementally trains the model with the input from the human workforce. We highlight potential biases that are likely to be propagated from different perspectives of human moderators who provide final decisions and labeling for re-training a classification model. To mitigate those biases, we propose an image annotation interface to develop an explainable dataset and the system that reflects various stakeholders' perspectives in understanding the 3D content. We conclude with initial recommendations for metadata design to (1) require consent and (2) inform previously unaware users of consent for publicizing the content which might invade copyright or privacy.

2. Algorithmic Content Moderation

Manual moderation relying on a few trusted human workforce and voluntary reports has been common solutions to review shared contents. Unfortunately, it becomes increasingly difficult to meet the demands of growing volumes of users and user-created content [12]. Algorithmic content moderation has taken an important place in popular social media platforms to prevent various sensitive content in real-time, including graphic violence, sexual abuse, harassment, and more. As with other media posts, 3D contents available online appear as web documents that consist of images and text. For example, to attract audiences and help others understand the design project, creators in Thingiverse voluntarily include various information such as written descriptions of the model, tags, as well as photos of a 3D printed design; thus, 3D content can provide us an ample opportunity to employ the existing text and image based moderation schemes.

Among various text-based solutions, sentiment analysis is one traditionally popular approach that categorizes input text into either two or more categories: positive and negative, or more detailed *n*-point scales (e.g., highly positive, positive, neutral, negative, highly negative) [5, 6]. Moderators can consider categorization results in deciding whether the content is offensive or discriminatory [13]. Various classifiers, such as Logistic Regression Model, Support Vector Machine, and random forest, are actively used in detecting misogynistic posts on Twitter (e.g., [7, 8]). Jigsaw and Google's Counter Abuse Technology suggested Perspective API [14] provide a score on how *toxic* (i.e., rude, disrespectful, or unreasonable) the text comment is, using a machine learning (ML) model that was trained by people's rating of internet comments.

With the rapid improvement of Computer Vision (CV) technologies with machine learning, several image datasets (e.g., NudeNet Classifier dataset[15]) and moderation APIs enable developers to apply these ready-to-use mechanisms to their applications. For example, Microsoft Azure Content Moderator [1] classifies adult images into several categories, such as explicitly sexual in nature, sexually suggestive, or gory. DeepAI's Nudity Detection API [2] enables automatic detection of adult images and adult videos. Amazon Rekognition content moderation [3] detects inappropriate or offensive features in images and provides detected labels and prediction probabilities. However, many off-the-shelf services and APIs are often obscured, because it is hard for users to expect that the models are trained with fair ground-truths that can offer reliable results to various stakeholders with different cultural or social backgrounds without any biases, which we will discuss more in a detailed way in the following section.

2.1. Challenges in Moderating 3D Content

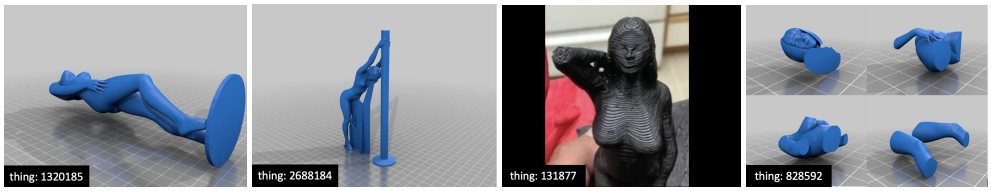
As we noted earlier, 3D contents appear as web documents that consist of text descriptions, auto-generated preview images, and user-uploaded images to help others comprehend the content at a glance. Although it is technically possible to utilize existing text and image based moderation schemes, 3D models have unique features that make it hard to directly adopt the existing CV techniques to their rendered images or photos.

2.1.1. 3D specific features that hamper the use of existing CV techniques

We identified four characteristics that make sensitive elements undetectable by the existing algorithms.

Challenge 1. Difficulties in Locating Features from Images of the Current Placement. Thingiverse automatically generates rendered images of the 3D model when a 3D file is uploaded, and this is used as a representative image if the designer does not provide any photos of real 3D prints. In many cases, these files are placed in the best orientation that guarantees print-success in FDM (Fused Deposition Modeling) printers, aligning the design to minimize overhangs. As the preview is taken in a fixed angle, so it might not be in a *perfect* angle that shows the main part of the model thoroughly (e.g., Fig 1(a)). It hinders the existing image-based APIs from accurate detection of sensitivity in the preview images, because sensitive parts might not be visible.

Challenge 2. Support Structure that Occludes the Features. Following the model alignment strategy of FDM printing, designers often include a custom support structure to prevent overhangs and to avoid printing failures and deterring surface textures with



(a) Rotated model (b) Support structure (c) Texture on surface (d) Divided into parts

Figure 1: Example images for the mainly 4 characteristics that make it hard to use the existing CV techniques; each thing is reachable using its unique ID through the url of https://thingiverse.com/thing:ID

auto-generated supports from slicers (i.e., 3D model compiler) such as Cura [16]. These special structures easily occlude the design's significant features (e.g., Fig 1(b)). Since the model is partly or completely occluded, the existing CV techniques barely detect sensitivity of the design.

Challenge 3. Texture and Colors. Current 3D printing technologies enable users to use various print settings and other postprocessing techniques. Accordingly, the printed model may present unique appearances compared to general real-world entities. Often the model is single-colored and can have a unique texture such as linear lines on the surface (e.g., Fig 1(c)) due to the nature of 3D printing mechanisms of accumulating materials layer-by-layer, which might let the existing CV algorithms overlook the features.

Challenge 4. Models Separated into Parts for Printing. As one common 3D printing strategy to minimize printing failures from a complex 3D designs such as a human body, many designers divide their models into several parts to ease the printing process, and let users post-assemble as shown in Fig 1(d). In this case, it is hard for the existing CV techniques to get the whole assembled model, resulting in a failure to recognize its sensitivity.

3. Transparency and Fairness Issues in Content Moderation

3.1. Transparency: Black Box that Lacks Explanation

Content moderation has long been controversial due to its non-transparent and secretive process [17], resulting from lacking explanations for community members about how the algorithm works. To meet the growing demands for transparent and accountable moderation practice as well as to elevate public trust, recently, popular social media platforms have begun to dedicate their efforts to make their moderation process more obvious and candid [17, 18, 19, 20]. As a reasonable starting point, those services provided detailed terms and policies (e.g., Facebook's Community Standards [21]) describing the bounds of acceptable behaviors on the platform [17]. In 2018, as a collective effort, researchers and practitioners proposed the Santa Clara Principles on Transparency and Accountability in Content Moderation (SCP) [22]. SCP suggests one requirement that social media platforms should provide detailed guidance to the members about which content and behaviors are discouraged, including examples of permissible and impermissi-

ble content, as well as an explanation of how automated tools are used across each category of content. It also recommends for content moderators to give users a *rationale* for content removal to assure about what happens behind the content moderation.

Making the moderation process transparent and explainable is crucial to the success of the community [23], in order not only to maintain its current scale but also to invite new users, because it may affect users' subsequent behaviors. For example, given no explanation about the content removal, users are less likely to upload new posts in the future or leave the community, because they may believe that their content was treated unfairly thus get frustrated owing to an absence of communication [24]. Reddit [25], which is one of the most popular social media, has equipped volunteer-based moderation schemes resulting in the removal of almost one fifth of all posts every day [26] due to violation of their community policy [27] (e.g., Rule 4: *Do not post or encourage the posting of sexual or suggestive content involving minors.*) or individual rules of the subreddits (i.e., subcommunity of Reddit that has a specific individual topic) according to their own objectives (e.g., One of the rules in 3D printing subreddit: "Any device/design/instructions which are *intended* injure people or damage property will be removed."). Users being aware of community guidelines or receiving explanations for content removal are more likely to perceive that the removal was fair [24] and showcase more positive behaviors in the future. As many social platforms including 3D open communities such as Thingiverse highly rely on voluntary posting of the user-created content [28], the role of a transparent system in content moderation becomes more significant in maintaining the communities themselves.

Even if many existing social media platforms have their full gears to implement artificial intelligence (AI) in content moderation, it has long been in the black box [23], thus not understandable for users due to the complexity of the ML model. To address the issue of the uninterpretable model that hinders the users from understanding how it works, researchers shed lights on the blind spot by studying various techniques to make the model *explainable* (e.g., [29, 30, 31]). Explainability has been on the rise to be an effective way of enhancing transparency of ML models [32]. In order to secure explainability, the system must enable stakeholders to understand the high-level concepts of the model, the reasoning used by the model, and the model's resulting behavior [33]. For example, as shown in the Fairness, Accountability, and Transparency (FAT) model, supporting users to know which variables are important in the prediction and how they will be combined is one powerful way to enable them to understand and finally trust the decision made by the model [34].

3.2. Fairness: Implicit Bias and Inclusivity Issues

People often overlook fairness of the moderation algorithm and tend to believe that the systems automatically make unbiased decisions [35]. In fact, the human adjudication of user-generated content has been occurred in secret and for relatively low wages by unidentified moderators [36]. In some platforms, users are even unable to know the presence of moderators or who they are [37], and thus it is hard for them to know what potential bias, owing to different reasoning processes, has been injected into the moderation procedure. For example, there have been worldwide actions that strongly criticize the sexualization of women's bodies without inclusive inference (e.g., 'My Breasts Are Not Obscene' protest by the global feminist group Femen [38] to denounce a museum's censor-

ship of nudity.). Similarly, Facebook's automatic turning down of postings and selfies that include women's topless photo by tagging them as *Sexual/Porn* ignited 'My Body is not Porn' movement [39, 40]. The different points of view in perceiving and reasoning towards the same piece of work makes it yet hard to decide the absolute sensitivity. It is nearly impossible that the sole group of users represent all, therefore, it is difficult for users to expect a *ground-truth* in the decision-making process, and trust the result while believing experts made the final decisions based on thoughtful consideration with an unbiased rationale.

Subsequently, many studies (e.g., [41, 42]) have explored potential risks of algorithmic decision-making that are potentially biased and discriminatory to a certain group of people such as underrepresented groups of gender, race, disability. Classifier has been one common approach in content moderation, but developing a perfectly fair set of classifiers in content moderation is complex compared to those in common recommendation or ranking systems, as classifiers tend to inevitably embed a *preference* to the certain group over others to decide whether the content is offensive or not [17].

3.3. Transparency & Fairness Issue in 3D Content Moderation

Through a text feature based classification, we identified there are three main categories of sensitive 3D content: (1) sexual/suggestive, (2) dangerous weaponry, and (3) drug/smoke. Due to the capability of unlimited replication and reproduction in 3D printing, unawareness of these 3D contents could be crucial. We noticed that Thingiverse limits access to *some* of sensitive things that are currently labeled as NSFW (Not Safe for Work) by replacing their thumbnail images with the black warning images. It is a secretive process because there are no clear rationale or explanations offered to users behind this process. Therefore, users cannot expect whether Thingiverse operates based on an unbiased and fair set of rules.

While the steep acceleration of increments of 3D models [43] is making automatic detection of sensitive 3D content imperative, moderating 3D content also faces fairness issues and users are suffering from lacking explanations. We need to take our account into various stakeholders' points of view that affect their decision on potentially sensitive 3D content, as well as further discussions to mitigate bias and discrimination of the algorithmic decision-making system. Here we propose an explainable human-in-the-loop 3D content moderation system to enable various users who have distinct rules to participate in calibrating algorithmic decisions to decrease bias or discrimination of the algorithm itself. Although we focus on specific issues in shared 3D content online, our proposed pipeline generally applies to advancing a semi-automatic process toward an explainable and fair content moderation for all.

4. Towards Explainable 3D Moderation System

A potential solution to examine 3D contents' sensitivity with fairness is employing the human workforce with ample experiences in observing and perceiving with various perspectives. We suggest a human-in-the-loop pipeline, based on the idea of incremental learning [44] that the human workforce can collaborate with an intelligent system, concurrently classifying data input and annotate features with the explanation for the decision.

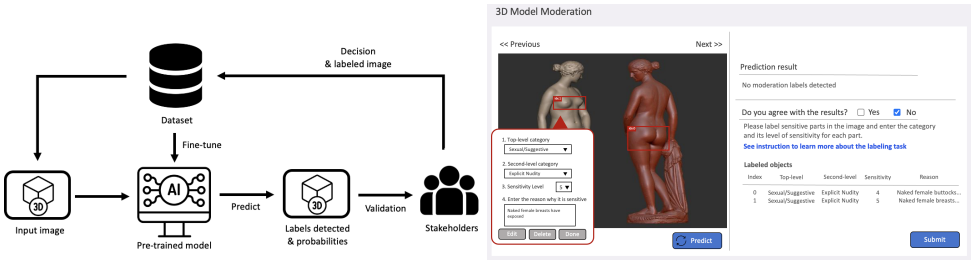
4.1. Building an Inclusive Moderation Process

Making decisions on the sensitivity of a 3D model can be subjective due to various factors such as cultural differences, the nature of the community, and the purpose of navigating 3D models. To reflect different angles in discerning the nature and intention of contents, we need to deliberate various interpretations taken from various groups of people. For example, there are lots of 3D printable replicas of artistic statues or Greek sculptures that are reconstructed by 3D scanning of the original in the museums [45]. Speculative K-12 teachers designing their STEAM education curriculum using 3D models are not likely to want any NSFW designs revealed to their search results. On the other hand, there are many activists and artists who may want to investigate the limitless potential of the technology, sharing a 3D scanned copy of the naked body of herself [46] or digitizing nude sculptures available in the museum to make the intellectual assets accessible to everyone, etc. The nude sculpture has been one popular form of artistic creation in history, and it is not simple to stigmatize these works as 'sensitive'. Everyone has their own right to 'leave the memory of self' in a digital form. Forcing to adapt a preset threshold of sensitivity and filter these wide array of user-created contents could unfairly treat one's creative freedom. As the extent that various stakeholders perceive the sensitivity could be distinct, our objective is to design an inclusive process in accepting and adopting the sensitivity.

4.2. Solution 1: Human-in-the-loop with Augmented Learning

Automated content moderation could help review of a vast amount of data and provide filtered cases for humans to support a decision-making process [24], if we well-echo diverse perspectives in understanding contents. In our proposal of the human-in-the-loop pipeline (Fig 2(a)), an input image dataset of 3D models will be used for the initial model training, then the result will be reviewed by multiple human moderators step by step. We trained the model with 1,077 things that are already labeled as NSFW by Thingiverse and 1,077 randomly selected non-NSFW things. All input images are simply categorized as NSFW or not, with no annotation for specific image features to provide the reasoning. Human moderators recruited from various groups of people now review the classification results whether they agree. They are asked to annotate image segments using a bounding box where they referred to make the final decision with the category. At the same time, they provide the rough level of how much the part affected the entire sensitivity and a written rationale for the decision. These features will enhance the data quality so to be used to fine-tune the model with the weighted score, thus the model becomes able to recognize previously unknown sensitive models based on the similarity and now can *explain* sensitive features.

When two different groups of people with different standards do not agree on the same model's classification results, the model uses their decision, annotated features, and levels of sensitivity to differentiate the extent of perceived sensitivity and reflect to the different threshold. For example, one moderator thinks that the model is sensitive while the other does not, the model will have a higher threshold in categorizing the content. Different decisions on the same model finally could be brought to the table for further discussion if needed, for example, to regulate policy guidelines, or used as search criteria for other community users who have similar goals in viewing and unlocking analogous 3D



(a) Human-in-the-loop pipeline (b) User interface mockup

Figure 2: (a) Overview of the human-in-the-loop pipeline powered by human moderators to acknowledge various perceptions of sensitivity and (b) an user interface mockup for the moderators to validate prediction results and provide annotations regarding their rationale, thus to augment the model.

contents. To summarize, one iteration contains the following steps:

1. The pre-trained model presents prediction results.
2. The human moderator can enter disagreement/agreement with the results and annotate sensitive parts with a sensitivity level and a decision rationale.
3. The annotated image is used to fine-tune the model.
4. If the decision for the image is different from other moderators, annotations and sensitivity levels are used to set the different threshold.

We elaborate more on feedback from the moderators by showing three possible scenarios: (1) the moderator's agreement with the prediction results, (2) sensitive parts not detected, and (3) false-classification of insensitive features sensitive.

Case 1. Agreement with the Prediction Result In case that the moderators agree with the decision, they can either finalize it or reject the classification, by selecting provided top-level categories (e.g., sexual/suggestive, weaponry, drug/smoke) and second-level categories (e.g., under sexual/suggestive, explicit nudity, adult toys, sexual activity, etc.). We currently refer to a two-level hierarchical taxonomy of Amazon Rekognition to label categories of inappropriate or offensive content.

Case 2. Sensitive Parts Ignored by the Algorithm Another possible case is that the specific feature in the image that the moderator perceives as sensitive is missing in the detection results. In this case, human moderators can label that part and provide rationales using *enter the level of sensitivity* field from 1 (slightly sensitive) to 5 (highly sensitive), how each specific part affects the entire sensitivity of the model.

Case 3. False Negative It is also possible that some parts detected by the model are not sensitive for the moderator due to the higher tolerance to sensitivity. The moderator can either submit the disagreement or provide more detailed feedback by excluding specific results.

Different *degrees* of sensitivity perception from various stakeholders can reflect distinct points of view, which may manifest fairness in algorithmic moderation through multiple iterations of this process. In our interface for the end-users that assists searching 3D designs, we let users set their desired threshold. For those who might find it difficult to decide a threshold that perfectly fits their need, we

show several random example images that have detected sensitive labels with the corresponding threshold. This pipeline also helps obtain the explainable moderation algorithm. Our model can help users understand the rationales of the model by locating detected features/prediction probabilities in the image and providing written descriptions that the moderators entered for data classification.

4.3. Solution 2: New Metadata Design to Avoid Auto-Filtering

Another potential problem in open 3D communities is copyright or privacy-invasive contents that are immediately marked as NSFW by Thingiverse indicating they are *inappropriate*. Currently, Thingiverse lacks notification and explanation for content removal, while a majority of them might invade copyrights. Its obscurity results in a negative impact on the user's future behaviors. For example, creators are frustrated at the un-notified removal of their content thus decided to quit their membership (e.g., [47]), which might not happen if they saw an informative alert when they post the content. Along with advanced 3D scanning technologies [48], many creators are actively sharing 3D scanned models (e.g., As of December 2020, Thingiverse has 1150 things that tagged with '3D_scan' and 308 things with the tag '3D_scanning'). With arising concerns over possible privacy invasion in sensitive 3D designs, what caught our attention is 3D scanned replicas of human bodies. Many of them do not include an explicit description of whether the creator received the consent from the subject (e.g., [49, 50]). Some designers quoted the subject's permission, for example, one creator describes that the subject, Nova, has agreed to share her scanned body on Thingiverse [51]. Still, this process relies on the users' voluntary action given no official guidelines, resulting in a lack of awareness that the users must be granted the consent to upload possibly privacy-invasive contents at the time of posting those content in public spaces regardless of the commercial purpose. Without explicit consent, the content is very likely to be auto-filtered by Thingiverse, which decreases fairness by hampering artistic/creative freedom. To iron out a better content-sharing environment in the these open communities, redesigning of metadata must be considered and adapted by system admins that invoke responsible actions. For example, providing a checkbox that asks *"If the design is made of 3D scanned human subject, I got an agreement from the subject"* can inform previously unaware users about the need for permission to post potentially privacy-breaching contents. Including the subject's consent can also protect creative freedom from auto-filtering, by adding that the content is not breaching copyright or privacy and can be shared in the public spaces. In addition, it can enable users to understand that an absence of consent could be the reason for filtering.

5. Conclusion

As an inclusive process to develop transparent and fair moderation procedure in 3D printing communities, our study proposes to build an explainable human-in-the-loop pipeline. We aim to employ diverse group of human moderators to collect their rationales, which can be used to enhance the model's incremental learning. Our objective is not to censor 3D content but to build a pleasant 3D printing community for all, by safeguarding search as well as guaranteeing creative freedom, through the pipeline and new metadata design that has potential to minimize issues related with privacy or copyright.

References

[1] MicroSoft, Adult, racy, gory content: Azure cognitive services, https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-detecting-adult-content, 2020. (Accessed on 05/21/2020).

[2] DeepAI, Nudity detection api, https://deepai.org/machine-learning-model/nsfw-detector, 2020. (Accessed on 05/21/2020).

[3] A. W. Services, Amazon rekognition content moderation, 2020. URL: https://docs.aws.amazon.com/rekognition/latest/dg/moderation.html, (Accessed on 12/20/2020).

[4] Thingiverse, 2008. URL: https://www.thingiverse.com/.

[5] R. Prabowo, M. Thelwall, Sentiment analysis: A combined approach, Journal of Informetrics 3 (2009) 143–157.

[6] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining., in: Lrec, volume 10, 2010, pp. 2200–2204.

[7] P. Saha, B. Mathew, P. Goyal, A. Mukherjee, Hateminers: Detecting hate speech against women, arXiv preprint arXiv:1812.06700 (2018).

[8] R. Ahluwalia, H. Soni, E. Callow, A. Nascimento, M. De Cock, Detecting hate speech against women in english tweets 330 (2018).

[9] S. Minaee, H. Pathak, T. Crook, Machine learning powered content moderation: Computer vision applications at expedia, Expedia Group Technology (2019).

[10] A. Kumar, N. K. Kumar, M. Shivaram, S. G. Jadhav, C.-S. Li, S. Mahadik, Image content moderation, 2020. US Patent 10,726,308.

[11] E. Llansó, J. Van Hoboken, P. Leerssen, J. Harambam, Artificial intelligence, content moderation, and freedom of expression (2020).

[12] D. Hettiachchi, J. Goncalves, Towards effective crowd-powered online content moderation, in: Proceedings of the 31st Australian Conference on Human-Computer-Interaction, OZCHI'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 342–346. URL: https://doi.org/10.1145/3369457.3369491. doi:10.1145/3369457.3369491.

[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2011) 267–307. URL: https://doi.org/10.1162/COLI_a_00049. doi:10.1162/COLI_a_00049.

[14] Jigsaw, G. C. A. T. team, Perspective api, 2017. URL: http://perspectiveapi.com/.

[15] bedigaadu, Nudenet classifier dataset, 2019. URL: https://archive.org/details/NudeNet_classifier_dataset_v1.

[16] Ultimaker cura, 2020. URL: https://ultimaker.com/software/ultimaker-cura.

[17] 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.

[18] N. Granados, A. Gupta, Transparency strategy: Competing with information in a digital world, MIS quarterly (2013) 637–641.

[19] K. Leetaru, Without transparency, democracy dies in the darkness of social media, 2018. URL: https://www.forbes.com/sites/kalevleetaru/2018/01/25/without-transparency-democracy-dies-in-the-darkness-of-social-media/?sh=479d8b527221#694732567221, (Accessed on 12/16/2020).

[20] M. MacCarthy, Transparency requirements for digital social media platforms: Recommendations for policy makers and industry, Transatlantic Working Group (2020).

[21] Facebook, Community standards, https://www.facebook.com/communitystandards/, 2020. (Accessed on 12/12/2020).

[22] U. o. S. C. U. Queensland University of Technology (QUT), E. F. F. (EFF), The santa clara principles on transparency and accountability in content moderation, 2018. URL: https://santaclaraprinciples.org/, (Accessed on 12/16/2020).

[23] P. Juneja, D. Rama Subramanian, T. Mitra, Through the looking glass: Study of transparency in reddit's moderation practices, Proceedings of the ACM on Human-Computer Interaction 4 (2020) 1–35.

[24] S. Jhaver, D. S. Appling, E. Gilbert, A. Bruckman, "did you suspect the post would be removed?" understanding user reactions to content removals on reddit, Proceedings of the ACM on human-computer interaction 3 (2019) 1–33.

[25] Reddit, https://www.reddit.com/, 2005.

[26] S. Jhaver, D. S. Appling, E. Gilbert, A. Bruckman, Did you suspect the post would be removed?": User reactions to content removals on reddit, Proceedings of the ACM on Human-Computer Interaction 2 (2018).

[27] Reddit, Reddit content policy, 2020. URL: https://www.redditinc.com/policies/content-policy, (Accessed on 12/12/2020).

[28] F. Çömlekçi, Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media, Communication Today 10 (2019) 165–166.

[29] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, et al., Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model, The Annals of Applied Statistics 9 (2015) 1350–1371.

[30] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, P. MacNeille, A bayesian framework for learning rule sets for interpretable classification, The Journal of Machine Learning Research 18 (2017) 2357–2393.

[31] A. A. Freitas, Comprehensible classification models: a position paper, ACM SIGKDD explorations newsletter 15 (2014) 1–10.

[32] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, P. Vinck, Fair, transparent, and accountable algorithmic decision-making processes, Philosophy & Technology 31 (2018) 611–627.

[33] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, P. Eckersley, Explainable machine learning in deployment, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 648–657.

[34] H. Lakkaraju, S. H. Bach, J. Leskovec, Interpretable decision sets: A joint framework for description and prediction, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1675–1684.

[35] S. Garfinkel, J. Matthews, S. S. Shapiro, J. M. Smith, Toward algorithmic transparency and accountability, Communications of the ACM 60 (2017) 5–5.

[36] S. T. Roberts, Behind the screen: Content moderation in the shadows of social media, Yale University Press, 2019.

[37] S. T. Roberts, Commercial content moderation: Digital laborers' dirty work (2016).

[38] S. Cascone, Topless feminist protestors hit the musee d'orsay after the museum tried to bar a visitor for wearing a low-cut dress, 2020. URL: https://news.artnet.com/art-world/femen-stage-protest-musee-dorsay-1908260, (Accessed on 12/22/2020).

[39] B. Korea-savvy, Activists claim "my body is not porn!", http://koreabizwire.com/activists-claim-my-body-is-not-porn/119529, 2018. (Accessed on 09/08/2020).

[40] My body is not your porn, 2020. URL: https://www.facebook.com/pages/category/Community/My-Body-Is-Not-Your-Porn-106365187645422/.

[41] S. L. Blodgett, L. Green, B. O'Connor, Demographic dialectal variation in social media: A case study of african-american english, arXiv preprint arXiv:1608.08868 (2016).

[42] 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.

[43] B. Wire, Makerbot thingiverse celebrates 10 years of 3d printed things, 2018. URL: https://financialpost.com/pmn/press-releases-pmn/business-wire-news-releases-pmn/makerbot-thingiverse-celebrates-10-years-of-3d-printed-things, (Accessed on 12/10/2020).

[44] M. Längkvist, M. Alirezaie, A. Kiselev, A. Loutfi, Interactive learning with convolutional neural networks for image labeling, in: IJCAI 2016, 2016.

[45] C. Marshall, 3d scans of 7,500 famous sculptures, statues & artworks: Download & 3d print rodin's thinker, michelangelo's david & more, 2017. URL: https://www.openculture.com/2017/08/3d-scans-of-7500-famous-sculptures-statues-artworks-download-3d-print-rodins-thinker-michelangelos-david-more.html, (Accessed on 12/23/2020).

[46] B. Mufson, Art made from human body scans | gif six-pack, 2016. URL: https://www.vice.com/en/article/nz4kq7/3D-scanning-gifs, (Accessed on 12/23/2020).

[47] VidovicArts, I'm quitting thingiverse, 2020. URL: https://www.youtube.com/watch?v=UPRCE8FsSak.

[48] All3DP, 2020 best 3d scanners (december), 2020. URL: https://all3dp.com/1/best-3d-scanner-diy-handheld-app-software/, (Accessed on 12/23/2020).

[49] Tob1112, 3d body scan amber, 2015. URL: https://www.thingiverse.com/thing:1052758.

[50] ThreeForm, Mel - "column 2" pose, 2017. URL: https://www.thingiverse.com/thing:2688184.

[51] ThreeForm, Nova - "pose 3", 2017. URL: https://www.thingiverse.com/thing:2461567.