Filtering harm: the politics of toxicity in content moderation infrastructures

Journal:	*New Media and Society*
Manuscript ID	NMS-20-0648
Manuscript Type:	Original Manuscript
Keywords:	Content moderation, toxic discourse, pollution behaviour, Perspective API, Opt Out, online communities, boundary work, filtering, dirt, NLP
Abstract:	This article conceptualizes content moderation infrastructures as filtering systems set in place to preserve the 'health' of online communities. We argue the notions of 'dirt' and 'pollution' are valuable for understanding the logics of these filtering systems, and the deeper structural they are embedded within. To substantiate this, we present to content moderation initiatives, Perspective API and Opt Out, to show how toxic content is defined, detected and negotiated. These two cases illuminate the multiple meanings 'toxicity' has come to bear in content moderation, the cultural filtering work it has come to do, and the complex politics to which it has given rise. And they show that content moderation's historical reliance on static categories – and those categories' ongoing entanglements with social systems of racism and patriarchy – embeds content moderation systems in structural ideas about community health and harm that often end up amplifying inequalities.

Filtering harm: the politics of toxicity in content moderation infrastructures

Introduction

The advent of large-scale social media platforms has given rise to widespread online abuse, harassment and exploitation. 'Toxic' content has become an established way of understanding and describing the results of this behaviour. Yet, while the concept is becoming naturalised as a way of expressing harm, its uses are also mobilised to very different ends. We wish to use this article to explore the role the notion of 'toxic' has come to play in online content moderation, and what the politics of this discourse implies.

Our overarching aim is to ask how content moderation infrastructures define toxic content and what the politics of this definition are. We argue that content moderation's historical reliance on static categories – and those categories' ongoing entanglements with social systems of racism and patriarchy – embeds the infrastructures in structures that risk reproducing inequalities. We use the content moderation initiatives Perspective API and Opt Out as examples to explore content moderation infrastructures define, identify and handle potentially harmful content. We have chosen these two examples to illustrate the difference between top-down and bottom-up approaches. The two cases do not provide an exhaustive mapping of content moderation, but rather illuminate the multiple meanings 'toxicity' has come to bear in content moderation, the cultural filtering work it has come to do, and the complex politics to which it has given rise. Moreover the examples help us foreground the challenges inherent in attempts to automate and scale content moderation as well as circle in on the two fundamental questions of content moderation, namely: whom are content moderation rules for? And who gets to define and enforce them?

While other scholars have offered necessary and important critiques of content moderation as a form of governance (Gillespie, 2018; Klonick, 2018; Gorwa, Binns and Katzenbach, 2020), human-machine processes (Ruckenstein and Turunen, 2019) and digital labour (Roberts, 2015; Nakamura, 2016), we believe that the discourse of toxicity in content moderation infrastructures warrants further scrutiny. We show that scholarly works that draw on and develop pollution and discard theory help us to understand this discourse and its politics and point to new possible and desirable avenues of research in content moderation studies. Relying on theories of social pollution developed in anthropology (Douglas, 1966), as well as on more recent work on dirt and toxicity in the emerging field of discard studies (Liboiron, Tironi and Calvillo, 2018; Lepawsky, 2019b), we argue that content moderation should not be seen merely as a question of negative removal of toxic content, but also as a productive 're-ordering of our environment' through practices of classification and purification (Douglas, 1966: 2).

Theoretical approaches to content moderation

Content moderation infrastructures have been around for as long as online communities have existed (Roberts, 2019). With time three clearly discernible strands of research literature on content moderation technologies have emerged, which differ in terms of their disciplinary embeddings and associated research interests: a) media, communication and information studies, b) law, and c) computational theories regarding natural language processing (NLP) and computer vision. Combined, these three strands have drawn up important contours of content moderation as a question of mediation; labour; legislation; and technology.

The application of content moderation technologies as a question of mediation has been studied through interdisciplinary work at the intersection of communication theory, information studies and digital platform studies. This strand of research centres on how content moderation technologies are deployed as self-regulatory apparatuses, even platform commodities, to prevent and identify problematic content and behaviour with a view to improving user experience and negotiating the political role of platforms (Gillespie, 2018; Roberts, 2019). A key idea is that because of content moderation's ability to discreetly, even imperceptibly, remove potentially offensive content, such infrastructures are 'in many ways, the commodity that platforms offer' and thus resource platforms such as Facebook and Instagram offers (Gillespie, 2018). Other strands of communication and digital media-related literature examine the effects of content moderation on the interactions of users whose content falls into the category of the 'problematic', such as 'pro-ana' communities (Cobb, 2017; Gerrard, 2018), as well as the impact of content moderation on marginalised groups, for instance queer women (Duguay, Burgess and Suzor, 2018). Yet other strands within this literature examine content moderation from the perspective of moderator engagement, and how this engagement in turn shapes communities (Nakamura, 2016; Dosono and Semaan, 2019; Gibson, 2019; Seering *et al.*, 2019; Matias, 2019) and performs a form of civic labour (Matias, 2016).

If media and communication studies approach content moderation as a problem of power, governance and communication, legal-political approaches tend to focus more on fundamental legal-political categories such as regulation (Grimmelmann *et al.*, 2015), free speech (Klonick, 2018), censorship (Deibert, 2009) and human rights (Jørgensen and Zuleta, 2020). Legal studies of content moderation thus focus on different sets of codes within content moderation assemblages, often with the aim of optimising legal frameworks to ensure that content moderation filtering mechanisms work in compliance with other concerns such as human rights.

The third body of research, which is associated with the optimisation of the technologies themselves, has developed in response to the demand for standardized and scalable content moderation systems as the amount of abusive language online has escalated, and it has primarily focused on digital systems and devices that allow platforms and users to detect, analyse and manage content (e.g. the Perspective API by Jigsaw). The key aim in this body of work is to test the efficiency and accuracy of filters in detecting harmful content (Nobata *et al.*, 2016; Jaki *et al.*, 2019); testing their robustness (Hosseini et al., 2017), examining bias (Davidson et al., 2019; Jiang et al., 2019) and building typologies of abusive language (Davidson et al., 2017; Waseem et al., 2017). For instance, Waseem (2016) surveys the differences in content moderators' annotation practices based on their professional and life experiences, and theorises the implications of this for the development of corpora and annotation guidelines. This body of work divides into two strands that often complement each other. One strand emphasises the innovative potential of content moderation technologies themselves (Yang et al., 2019). The other strand seeks to uncover the complications and limitations of automating content moderation (Davidson et al., 2019). One strong line of argument in both strands concerns the difficulty of transferring humans' contextual knowledge to machinic operations. More recently, critiques have arisen that the data used to train content moderation tools in many cases represents normative language use, to the detriment of marginalised communities (Gomes et al., 2019; Sap et al., 2019). When we scale content moderation, we also run the risk of overgeneralising through misclassification, restricting users' freedom of expression, jeopardising users' reputations, misclassifying abusive conversations as ordinary conversations, maintaining the status quo of abusive online communities, and further punishing communities that are already marginalised (Gomes et al., 2019).

While these bodies of literature on content moderation technologies demonstrate significant scope and breadth, a deeper theorisation of content moderation's reliance on notions of toxicity and dirt has been absent, in both theoretical and empirical terms. This is surprising, since the industry's stated problem – that is, how to detect, contain, prevent and 'clean up' abusive behaviour – is symptomatic of much more pervasive scholarly and popular discourses of content moderation that rely heavily on metaphors of waste, discard, pollution, cleaning and maintenance. From documentaries depicting the working conditions of content moderators *The Cleaners* (dir. Lisa Molomot, 2018) to scholarly work on content moderation and common parlance in the content moderation industry, terms such as 'cleaning', 'scrubbing', 'janitors', 'detritus', 'refuse', 'dirt', 'health', 'sanitisation' and 'waste' are widespread (Roberts, 2015; Ruckenstein and Turunen, 2019; Parks 2019). Thus empirical studies and investigative journalism scrutinise how the content moderation industry is doing the 'dirty work' (Roberts, 2016) of 'cleaning up abusive content' (Roberts, 2015: 111), 'tackles all kinds of "mess" and "disorder"' (Ruckenstein and Turunen, 2019: 6), and provides a checkpoint for 'dirty stuff' that allows users to freely share and connect while at the same time 'disavowing that in many ways users want that sharing to be cleansed of the aspects of human social exchange they find abhorrent' (Gillespie, 2018). These discourses have in turn given rise to contestation of classification schemes that spawns moral debates over whether or not an expression is harmful. How and when, for instance, are images of nudity and porn harmful, to whom do they pose harm, and are they harmful because of the effects they might have on the viewers or on the nude models or performers (Stardust 2014; Breslow 2018; Bronstein, 2020)? Should images of war atrocities be considered harmful, and if so, are they harmful because they traumatise viewers, because they put combatants and civilians in danger, or because they can be misused (Sontag, 2002; Banchik, 2020; Saber, in press)? Disregarding the many fundamental uncertainties that beset digital material (Ekman *et al* 2017), many digital platforms offer content moderation technologies as simple and scalable solutions. Yet, while they often pride themselves of constantly improving their ability to 'clean up' toxic content through technological means, they also, as we show in this article, often end up reproducing racist and misogynist patterns.

Theorising content moderation as a problem of dirt

This article argues that content moderation studies must reckon with the inseparability of hygiene and pollution discourses and content moderation technologies as 'protective' filtering systems that reject and accept to ensure the 'health' of communities. To date we have very few theoretical tools to grasp the politics of content moderation discourses on toxicity and their entanglement in digital 'pollution behaviors' (Douglas, 1966). In this article, therefore, and following Josh Lepawsky's lead (2019a), we use Mary Douglas's anthropological framework of 'dirt' to conceptualise content moderation technologies as systems put in place to protect platforms and their communities against existential threats. Lepawsky points out that Douglas's work can help us to understand online communities as systems that must rid themselves of things in order to be able to constitute themselves as such, and content moderation technologies as the filters that help them do this work.

As Douglas notes in her now classic work on the ideas of purity and danger, 'dirt is the by-product of a systematic ordering and classification of matter, in so far as ordering involves rejecting inappropriate elements' (Douglas, 1966). For Douglas, 'no single item is dirty apart from [i.e. outside of] a particular system of classification in which it does not fit' (Douglas, 1966b: vii). Dirt depends on a system (Douglas, 1966b: 44) because it is not an independent, objective attribute of something, but a 'residual category [of things] rejected from our normal scheme of classifications' (Douglas, 1966b: 45). Dirt is a label for 'all events which blur, smudge, contradict or otherwise confuse accepted classifications' (Douglas, 1966b: 50), and importantly, it is a contextual term: 'what is clean in relation to one thing may be unclean in relation to another' (Douglas, 1966b: 10). Douglas illustrates these points with mundane examples: shoes, for instance, are not dirty in themselves, 'but it is dirty to place them on the dining table' (Douglas, 1966b: 44). Similarly, food is not necessarily dirty, 'but it is dirty to leave cooking utensils in the bedroom' (Douglas, 1966b: 37). She proposes that culture and its organisation are practised through rituals and habits that bind individuals to a group and establish group borders, which are maintained via structural distinctions such as purity and pollution, or dirtiness and cleanliness. Such distinctions might thus be described as a form of border hygiene that helps to maintain the integrity of identities, individual or collective.

In *Purity and Danger*, Douglas describes how societies maintain a sense of purity through the systematic separation and demarcation of those subjects or objects considered dirty or pollutant. The expulsion of the impure creates a sense of unity, as members cohere around a shared meaning (Douglas, 1966b: 2). Dirt is therefore something communities avoid in order to prevent the breakdown of meanings, and on a larger scale the breakdown of the community itself. It is no coincidence that waste management is often foregrounded as a civilising sign, or that the obstruction of waste management or mobilisation of waste is one of the most efficient means of protest (Foucault, 2006: 26; Moore, 2008). Similarly narratives of civilisation, progress and perfection also often refer to waste, hygiene and abject matter (Freud 1961, 47). While these strands of thought emphasises dirt avoidance, Douglas argues that eliminating actions are not negative processes of removal; rather, as she notes, dirt removal is a 'positive effort to organize the environment' (Douglas, 1966b; 2) of the community in which it takes place. Hence, Douglas argues:

> In chasing dirt, in papering, decorating, tidying, we are not governed by anxiety to escape disease, but are positively re-ordering our environment, making it conform to an idea. There is nothing fearful or unreasoning in our dirt-avoidance: it is a creative movement, an attempt to relate form to function, to make unity of experience. (Douglas, 1966b; 2)

Yet while many would consider themselves able to detect dirt fairly easy (dog pooh on one's shoe, vomit on one's shirt and dust in the corners), Douglas reminds us that detecting dirt is in fact anything but straightforward, since 'there is no such thing as absolute dirt: it exists in the eye of the beholder' (Douglas, 1966b; 2). Thus what appears as dirt to one beholder, might appear as a valuable resource by another. Dog poo looks very different to a biologist interested in bacterial growth and a professor accidentally stepping in a turd on their way to work. And while pork is a delicacy to some, it is an inherently dirty food to others.

Douglas's insights help us identify the complexity of content moderation technologies and their classification schemes of harmful content: firstly, content moderation requires constant creative efforts to classify, detect and reorganize content online expressed in every last bit of its mechanisms from computational models to conceptual frameworks and from organizational systems to manual labour. Within content moderation these elements become different mechanisms that operate productively and positively to reorganise online environments through removal and relegation. In both Facebook's friendly space of holidays images and shared special moments between friends, families and lovers as well as Instagram's carefully curated aesthetic of attractive bodies, delicious foods and urban and rural landscapes, content moderators are hard at work, removing war atrocities, child pornography and, in the case of TikTok, even bodies classified as "ugly, poor or disabled users" (Biddle et at, 2020). In all these cases, removal is not only a negative act, but also part of a productive process embedded in complex community formations and co-optations (Bucher 2020).

Secondly, Douglas's framework allows us to see the cultural contingency of content moderation: what constitutes an 'offence against order' often gives rise to intense negotiations of between coders (Waseem and Hovy, 2016) and designers of data (Waseem, 2016) often struggle to properly operationalise 'hate speech', 'toxicity' and 'offensive language' in automated detection. Further, as Waseem (2016) notes, even if what humans codify as dirt is highly subjective and thus indeterminate, machine learning technology ends up representing the classification schemes as truth. Algorithms thus identify the boundaries of the bodies of norms the annotators have encoded into the data and make it into universalized rules.

Stuart Hall's semiotic theory of cultural meaning-making, in which he draws on Mary Douglas, offers one explanation as to why the problem of indeterminacy continues to haunt content moderation practices (Hall, 1997). As Hall (1997: 236) notes, classifications are important to human beings because they are fundamental to meaning-making processes. Every culture has an order of classification built into it, and this seems to stabilise the culture.

Therefore, as Stuart Hall notes with reference to Douglas's work, it is also destabilising elements that cause most cultural concern.

> Mary Douglas argues that what really disturbs cultural order is when things turn up in the wrong category; or when things fail to fit any category – such as a substance like mercury, which is a metal but also a liquid, or a social group liked mixed-race *mulattoes* who are neither "white" nor "black" but float ambiguously in some unstable, dangerous, hybrid zone of indeterminacy in-between. Stable cultures require things to stay in their appointed place. (Hall, 1997: 236)

At the same time, as Hall's (1997) theory of encoding and decoding makes clear, an expression might be encoded with a specific intended reading by the sender, but it will then be interpreted, or decoded, by the reader from one of three positions: dominant, negotiated or oppositional. These moments of interpretation create a space of uncertainty and potential instability, where things that may have been encoded with one meaning by the sender can be decoded with a very different meaning by the receiver. Hence, someone could send a picture of a peach to indicate a desire to eat a piece of fruit, but the reader might decode the message as a sexually charged symbol of a pair of buttocks. Indeed, such an oppositional reading can give rise to subcultural communities, that stabilise their own meaning-making processes so that everyone in that community will understand the meaning of an eggplant.

Content moderation adds another level of complexity to Stuart Hall's model because it inserts a third layer of interpretation between the sender and receiver of messages. In content moderation the reader is not the intended recipient, but an intermediary that reads on behalf of the intended recipient in order to make a judgement regarding how the message will be received by the reader and hence whether it should be passed through or not. And yet, in doing this, content moderation systems also end up providing an absolute truth: the intermediary position becomes the final position, regardless of the recipient's perceptions or the sender's intention; instead an adjudication is made by an intermediary on behalf of the recipient. This makes content moderation a particularly challenging practice: when is porn pornographic, or abusive language abusive? And according to whom?

Understanding these complex meaning-making processes and different positions of power allows us to recognise that potentially problematic content might be flagged as problematic exactly because of its 'inability to be assimilated into existing socio-cultural categories and systems' (Rafi, 2015). In handling these types of expressions content moderation becomes a relative practice constantly oscillating among the meanings of content in specific contexts.

Cultural systems, moreover, can change quickly. What was once accepted practice, e.g. rape jokes, can suddenly be considered harmful and socially transgressive. Similarly, images that were once taboo in many communities, e.g. breastfeeding, could eventually become accepted in globalized public discourses. These cultural dynamics emphasise that when we talk about harmful content in content moderation, it is often less a question of the 'essence' of an expression and more a question of the properties that are attached to the content.

These dynamic complexities are the root cause of the problems encountered by standardized content moderation practices and technologies: how to assess whether a status is also a pornographic image, or whether the N-word is being used as a racist slur or a 'soul' word (Rahman, 2012)? Automated content moderation systems assign each bit of content a probability. Rarely will the probability be an absolute zero, reinforcing Douglas's assertion that there is no such thing as absolute dirt. In fact, many of the methods that seek to reduce the costs such systems impose on marginalised people specifically seek to lower the base of the probability (Liu and Avci, 2019) to a level where the system cannot distinguish it from 'clean' content. However, work that considers how the representation of gender in data sets influences classification systems has shown that automated systems go beyond simply representing inequities; rather, automated systems often end up amplifying them (Zhao et al., 2017).

As feminist, post- and anti-colonial scholars working on pollution theory emphasise, the capacity to define something as dirt – and thus as something to be rejected – is also a matter of power, bound up with structural issues of colonialism, patriarchy and race (Hall, 1997; Liboiron, Tironi and Calvillo, 2018). In his work on 'race' as a floating signifier, Hall links the classification of the dirty and the clean to racist logics of social purification:

> What you do with dirt in the bedroom is you cleanse it, you sweep it out, you restore the order, you police the boundaries, you know the hard and fixed boundaries between what belongs and what doesn't. Inside/outside. Cultured/uncivilised. Barbarous and cultivated, and so on. (Hall, 1999: 3)

Content moderation infrastructures subject marginalised communities to excessive policing, and disproportionately identify those communities' expressions as 'toxic'. In one article, Jessica Guynn thus recounts how Facebook deleted a user's comment after 15 minutes because it was deemed to be in violation of community standards on hate speech. The comment read:

'White men are so fragile and the mere presence of a black person challenges every single thing in them'. Indeed, the excessive policing of black users has given rise to the concept of 'getting zucked', and to new expressions such as 'wypipo' (used instead of 'white people') to outmanoeuvre content moderation infrastructures (Guynn, 2019). Sometimes the excessive filtering processes that reject black content as particularly polluting or dirt-like reproduce classic racialised logics. At other times, content moderation's static categories fail to capture the semantic richness, for instance, of African-American Vernacular English (AAVE). Most content moderation infrastructures thus work with the de facto assumption that the N-word is a negatively loaded word. As scholars show, data sets for abusive language and systems for detection encode structural biases against people who communicate in AAVE (Davidson *et al.*, 2017; Sap *et al.*, 2019). Such structural biases in part emerge because scholars working big data fail to capture what André Brock (2015), citing Lori Kendall, terms deep data i.e. methods that include deeper insights about the 'cultural, modal and social choices about technology use' we find in different cultural communities (Brock, 2015). For instance, in these systems the use of the N-word is a marker that typically automatically warrants penalisation of the user (Waseem, Thorne and Bingel, 2018). Yet, as Jacquelyn Rahman (Rahman, 2012) shows, for some African-American speakers the N-word represents a much more complex concept that, depending on the context of its usage, might 'convey a range of attitudinal stances related to its basic meaning, including solidarity, censure, and a proactive stance that seeks to bring about positive change'. Systems that centred their concept of the norm within AAVE-speaking communities would be unlikely to suggest that the N-word was unacceptable almost regardless of the context within which it was used.

To put it bluntly, then, many content moderation infrastructures reproduce the structural problems of respectability politics and its favouring of upper–middle class White ideals (Muhammad 2010; Pitcan *et al* 2018). As we show below, this is a problem content moderation infrastructures struggle with on a daily basis and also something that is being tackled differently. Below we offer two illustrative examples of different content moderation approaches to this problematic.

Handling toxicity online: two cases

Perspective AI

Perspective API was launched in 2017 by Jigsaw and Google's Counter-Abuse Technology team in a collaborative research project called Conversation-AI (Cellan-Jones 2017). The method employed by Perspective is to explore online discussions through experiments, models and research data, in order to create better governance tools and 'explore the strengths and weaknesses of [machine learning] as a tool for online discussion'. The API thus uses machine-learning models to score the 'toxicity' of an input text. Feeding a comment into the API delivers a score between zero and one; if the comment has a score of 0.9 or above, it is considered toxic. The API is 'trained by asking people to rate internet comments on a scale from "Very toxic" to "Very healthy" contribution', and the resulting conception of 'toxic' – defined by Perspective as 'a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion' – is thus partly informed by the people who trained the API. A wide range of online communities and publishers have implemented Perspective API's toxicity measurement system, including Wikipedia, which uses it on its editorial discussion pages; *The New York Times*, *The Guardian* and *El País*, which use it to parse their comments sections; and Reddit and Disqus, which use it to better understand the chilling effects of abuse in online discussions. The API outputs the scores in real time, so that publishers can integrate them into their websites to show toxicity ratings to commenters even during the typing.

In its brief self-presentation, Perspective API addresses the toxic as the flip side of the healthy. That is, it outlines a vision of a healthy conversation and its other. What, then, is unhealthy or toxic according to Perspective API? This question can partly be answered by looking at the training data released on GitHub, which includes sentences such as the following:

> Please check things out before blowing crap diarrhea out of your mouth. Also, its best to actually read what people write before jumping on an idiot bandwagon.
>
> You better Shut The Fuck Up and shove your Wikipedian Assholifity (yeah i made up that word what are gonna do sue me, huh bitch?) right up your sore-from-taking-it-up-the-ass asshole. I saw an attempt to give a nice little wikilink to a fun English usage, and you just fuck everything up. I'm gonna be bold, not like fucking pussies such as yourself.
>
> WHAT DID I DO??? FEMALE EJACULATION IS FUCKING GROSS!

These passages all reproduce classic gendered and sexualized 'dirty' taboos including feces, anal sex and female ejaculation (Segal 1999; Barcan 2005). But the training process is not only based on such 'found objects'. According to an interview with the team behind Perspective, Perspective also initially relied on crowdsourced abusive comments: 'In 2017, the team opened up the initial Perspective demo via public website as part of an alpha test, letting people type millions of vile, abusive comments into the site. It was kind of like Microsoft's infamous failed Tay chatbot experiment, except instead of tricking the bot into replying with racist tweets, Jigsaw used the crowdsourced virulence as training data to feed its models, helping to identify and categorize different types of online abuse.' (Marvin, 2019). Jigsaw then used this crowdsourced pool of aggressive expressions and combined it with training data sourced from Wikipedia and *New York Times* comments. Afterwards it added another crowdsourcing layer to its model, gathering 10 answers for each comment regarding whether it was toxic or not. The team members recruited raters through surveys to identify and categorise different types of online abuse within the training data. Yet, as the developers of Perspective also initially conceded, training data does not always lend itself to contextual analysis (Wakabayashi, 2017). One problem encountered in the deployment of the API is that it often flags unproblematic sentences as harmful. As librarian Jessamyn West discovered around the time the tool launched, the sentence 'I am a man' was considered only a little toxic, while 'I am a gay black woman' scored very high on the toxicity level. West found that 'I am a black trans woman with HIV' got a toxicity ranking of 0.77, 'I am a black sex worker' 0.89, and 'I am a porn performer' 0.80.

Figure 1. Jessamyn West Twitter

Such classifications reproduce structural perspectives on race, gender, sexuality and sex work as by definition "dirty" and taboo. In addition to these issues, the American writer David Auerbach also found flaws in relation to religion and persecution, showing that 'I fucking love you man. Happy birthday' scored 0.93 toxicity score, while 'Hitler's biggest mistake was not getting the job done was only deemed to be 0.06 toxic.

Figure 2. David Auerbach Twitter

Why did Perspective API attribute negative connotations to gender and sexuality and attribute positive – or indifferent – connotations to fascist/Nazi content? The reasons have to do with the ways in which we teach machines and they include both internal and external factors. The internal factors relate to the training data, and the external factors relate to word embeddings. Considering first the internal factors, Perspective API's ability to determine toxicity depends on the data used to train the model and the linguistic and lexical variety used within this labelled data. Machine learning methods then look for correlations within and between the data binned into different labels; thus if slurs or identity terms such as "Black", "gay", and "woman" occur at a higher rate in the positive, or toxic in terms of Perspective, labels, then those terms will have stronger correlations to positive classes, resulting in attributing more toxicity to identity sentences such as those documented by Jessamyn West. Conversely, content that employs 'civilized' language while arguing for positions that more profoundly disturb the social order is relatively less frequently occurring in the positive classes, and thus the model will not have learnt to strongly associate such language with toxicity in need of 'cleaning up'. Secondly the external factors: many models, and certainly some iterations of the Perspective API employ word embeddings to allow the model to draw on knowledge beyond what is contained in the data for the predictive task. In word-embedding models, machine-learning models assign use words as vectors to capture meaningful semantic relationships between corresponding words, e.g. 'Berlin' and 'Germany' and 'spaghetti' and 'food'. These models are typically trained automatically on large corpora of text, such as collections of Google Books or Wikipedia. However, as scholars have shown, word embeddings maintain stereotypical, and often negative and discriminatory positions, towards women and marginalised communities (Speer, 2017; Bolukbasi *et al.*, 2016; Zhao *et al.*, 2017; Caliskan, Bryson and Narayanan, 2017) whose are associated with harm, taboo and social pollution. Additionally, even embeddings that have undergone processes to address social biases against marginalised people maintain those biases through other representations of embeddings (Gonen and Goldberg, 2019). At the time of this writing, the phrase 'black queer women' scored 0.77 toxicity, while 'white men are' scores 0.25, and 'white straight men are' scored 0.50.

Opt Out

Our second case study is Opt Out, a Firefox browser extension founded by Theresa Ingram and developed by a team including Cheuk Ting Ho, Dr. Matteo Guzzo, Andrada Pumnea, Sophie Walker, Muaaz Saleem, Lucie Le Naour and Dr. Nicole Shephard. The extension, which was launched on 8 March 2020, focuses on misogyny, which it defines as 'any verbal, visual or physical harassment and abuse rooted in misogyny that is threatened, carried out and/or amplified online' (*How we define online misogyny*, no date). The mission of Opt Out is to address the torrent of misogyny online on an individual basis – removing misogyny from a single person's stream. This contrasts with Perspective API, which was developed as a tool for comment-enabled websites. Thus, while the Perspective API aims to develop a global understanding of toxicity, Opt Out aims to adjust to each individual's tolerance of misogyny, under a global understanding of what constitutes misogyny.

Figure. 3 NotNalise, Twitter

Foregrounding the cultural contingency of harmful expressions, Opt Out implements machine learning systems that are trained on multiple previously published data sets, with competing definitions and operationalisations of misogyny, thus countering essentializing tendencies. In addition to this, the Opt Out browser extension provides the user with the ability to distinguish between levels of severity in misogyny, thus allowing the user to identify the level that best aligns with their own understanding of misogyny – their own understanding of what constitutes dirt.

Figure. 4 (a) Flash_hoe and Figure. X (b) AquariaOfficial on Twitter

In Figure X, the tensions between how to understand the term "b*tch" in context are laid bare, as the model seeks to distinguish between reclaimed and pejorative uses of the word. Moreover, Figure X (a) appears both pejorative and misogynistic in the text-only reading the model performs, however considering the image of the object of the statement, a depiction of the respiratory virus COVID-19 hedges at least the pejorative nature of the text.
Meanwhile, the punishment of self-referencing identities other than white male are not punished, suggesting that the different understandings of misogyny within the datasets used may not penalise the mentioning of ones identities. As the developers of the extension have shared with the authors of this article, balancing the different understandings of misogyny and levels of allowable toxicity does not come without costs of its own. Machine learning systems rely on consistency regarding what constitutes a body of data to be able to identify decision boundaries between different bodies of labelled data. If these boundaries are not upheld, the machinery cannot minimise internal confusion about what is deemed to be harmful expressions. Hence, Opt Out's use of multiple data sets and competing definitions of misogyny introduces costs to the models consistency of the positive label, misogyny, because the different datasets have different annotation schemes and define misogyny distinctly ways. Mary Douglas's framework allows us to understand this noise to the learning algorithm less as a technical problem to be solved, and more as a fundamental cultural question of boundary setting and ambiguity-tolerance. Competing definitions of harmful expressions create competing signals for the model regarding which correlations to take advantage of or optimise for in its efforts to identify misogyny. Because the combined dataset has multiple competing definitions but a limited number of data points, the model built on it is limited in the number of nuances it can access of

any single understanding of dirt. The consequence is that a sparse modelling space is made even more sparse, as less data remains in the centre and more is pushed to the margins of the space. The data that is left in the centre tends to comprise highly normative understandings of what is and is not dirt.

The deeper politics of toxic discourses in content moderation

How might we understand the difficulties encountered by Perspective API and Opt Out as cultural-technical issues in the identification and reorganisation of dirt? And how are these challenges lodged in material-discursive matrices of toxicity and dirt? As Roopika Risam notes, toxic 'has become a cultural code word for the irritants and pollutants that disrupt our lived experience' (Risam, 2015). In this emerging 'toxic' discourse, Risam notes, the notion of the toxic 'is marshalled as the flip side of the healthy, the well'. Drawing on literary theorist Lawrence Buell's classic essay on the emergence of a particular toxic discourse led by environmentalist movements in the 20th century, Risam uses these observations to analyse the discursive power at stake in toxic discourses, for instance when white feminists determine fourth wave feminism as a toxic form of discourse. In the essay that informs Risam's article, Buell describes the key discursive motif of discourses on toxicity as 'a pervasive disenchantment from the illusion of the green oasis which is accompanied or precipitated by totalising images of a world without refuge from toxic penetration' (Buell, 2015: 648). Buell defines such toxic discourse as 'an interlocked set of topoi whose force derives partly from the anxieties of late industrial culture, partly from deeper-rooted habits of thought and expression' (Buell, 2015: 639). Drawing parallels to feminist positions, Risam points out that white feminists use the notion of toxic femininity to position 'women of color feminists as the disruptive bodies that transgress fictive, ideal feminist spaces on Twitter' and thus as bodies that should be filtered out and rejected to protect the 'health' of the feminist communities white feminists identify with (Risam, 2015).

We are inspired by Risam's identification of the politics of toxic discourses online and believe it is a helpful articulation of the complex issues of the very concept of toxicity in content moderation and the ways that 'structures define toxicity' (Liboiron, Tironi and Calvillo, 2018: 333; Thylstrup 2019). Moreover, Risam's analysis helps us recognize the political stakes in setting up the structural binaries of toxic vs. healthy and dirty vs. clean. This matters for how we think about content moderation: do we think of content moderation as a process that can help us return to a simpler time, when communities were less conflict-ridden and more like pastoral oases untouched by toxic penetration? Or as a form of complex cultural filtering that is deeply context-dependent in need of constant reflection of whose interests are represented, from which perspective and to what end. In our concluding remarks we will argue for the need to view it as the latter.

Concluding remarks

The theoretical framework of dirt and toxicity shows us that if our key ambition for content moderation is to ensure that online spaces are liveable for the communities that exist within them, then we have to accept that very few content moderation technologies can work as universal solutions. Moreover, we also have to abandon the idea of the 'sanitised' space (Byron 2019; Haison *et al* 2019). Indeed, as we have seen in this article, even though dirt has no essence, it is still circumscribed, and power struggles around dirt are often contestations of its borders. Queer theory provides a well-established theoretical articulation of the messiness and power dynamics of boundary work, offering insights into the violence wrought upon humans, cultures and environments by such divisions while recognising that these divisions also 'structure our thought in ways that are almost impossible to escape' (Schaffer, 2015), no matter how vehemently we may oppose them. As Guy Schaffer notes, camp theory is particularly attuned to the ways in which things and people can 'blur, transgress and cover in glitter boundaries between waste and not-waste', since 'camp offers a mode of celebrating, reappropriating, and rendering waste visible, without pretending that waste has stopped being waste' (Schaffer, 2015). Camp communities thus adopt and critically rework the value systems of dirt and filth by challenging and at the same time celebrating 'what is arguably the single most important and foundational cultural division, that between the dirty and the clean' (Hotz-Davies, Vogt and Bergmann, 2017: n.p). Relatedly, and inspired by Catherine D'Ignazio and Lauren Klein's (2020) *Data Feminism*, we argue that content moderation infrastructures need to continuously align with social and digital justice movements and never take categories of dirt for granted. As feminist and anti-racist social justice and digital labour movements have shown, the power to classify and detect of online toxicity is distributed unequally, producing a double problem: on one hand, marginalised communities experience the excessive policing of content; on the other hand, those communities carry the burden of being subjected to abusive language, while at the same time also struggling to set their own terms about what is 'dirty' and not (or even if this 'dirt' is enjoyable). Content moderation infrastructure should alleviate, not add to the stress of, marginalized communities – a good place to start examining whether and how they support that ambition is to reflect and evaluate the their conception and practice of purification.

Banchik, A. V. (2020) 'Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights–related content', *New Media & Society*. doi: 10.1177/1461444820912724.

Barcan, R (2005). 'Dirty Spaces: Communication and Contamination in Men's Public Toilets', *Journal of International Women's Studies*, 6(2): 7-23. Available at: http://vc.bridgew.edu/jiws/vol6/iss2/2

Biddle, S. *et al* (2020) 'Invisible Censorship: TikTok moderators to suppress posts by 'ugly' people and the poor to attract new users', *The Intercept*, March 16. https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/.

Bolukbasi, T. *et al.* (2016) 'Man is to computer programmer as woman is to homemaker? Debiasing word embeddings', in *Advances in Neural Information Processing Systems*.

Breslow, J. (2018) 'Moderating the 'worst of humanity': sexuality, witnessing, and the digital life of coloniality', *Porn Studies*, 5:3, 225-240, DOI: 10.1080/23268743.2018.1472034

Brock, A. (2015) 'Deeper data: a response to boyd and Crawford', *Media, Culture and Society*. doi: 10.1177/0163443715594105

Bronstein, C. 'Pornography, Trans Visibility, and the Demise of Tumblr'. *TSQ*, 7 (2): 240–254. doi: https://doi.org/10.1215/23289252-8143407

Bucher, T. (2020). Nothing to disconnect from? Being singular plural in an age of machine learning. *Media, Culture & Society*, 42(4), 610–617. https://doi.org/10.1177/0163443720914028.

Buell, L. (2015) '1. Toxic Discourse', in *Writing for an Endangered World*. doi: 10.4159/9780674029057-001.

Byron, P. (2019) 'How could you write your name below that?' The queer life and death of Tumblr, *Porn Studies*, 6(3): 336-349, DOI: 10.1080/23268743.2019.1613925

Caliskan, A., Bryson, J. J. and Narayanan, A. (2017) 'Semantics derived automatically from language corpora contain human-like biases', *Science*. doi: 10.1126/science.aal4230.

Cellan-Jones, Rory. (2017) 'Google's plan to make talk less toxic', *BBC News*, February 23. **https://www.bbc.com/news/technology-39063863.**

Cobb, G. (2017) '"This is not pro-ana": Denial and disguise in pro-anorexia online spaces', *Fat Studies*. doi: 10.1080/21604851.2017.1244801.

D'Ignazio, C. and Klein, L. F. (2020) *Data Feminism*, *Data Feminism*. doi: 10.7551/mitpress/11805.001.0001.

Davidson, T. *et al.* (2017) 'Automated hate speech detection and the problem of offensive language', in *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*.

Davidson, T., Bhattacharya, D., and Weber, I. (2019) 'Racial Bias in Hate Speech and Abusive Language Detection Datasets', in *Proceedings of the Third Workshop on Abusive Language Online*. doi: 10.18653/v1/W19-3504

Deibert, R. (2009) 'The geopolitics of internet control: Censorship, sovereignty, and cyberspace', in *The Routledge handbook of internet politics*.

Dosono, B. and Semaan, B. (2019) 'Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on reddit', in *Conference on Human Factors in Computing Systems - Proceedings*. doi: 10.1145/3290605.3300372.

Douglas, M. (1966) 'Purity-Danger - An Analysis of the Concepts of Pollution and Taboo', *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*.

Duguay, S., Burgess, J. and Suzor, N. (2018) 'Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine', *Convergence*. doi: 10.1177/1354856518781530

Ekman, U. Agostinho, D., Thylstrup, N. B. & Veel, K. (2017) The uncertainty of the uncertain image, *Digital Creativity*, 28:4, 255-264, DOI: 10.1080/14626268.2017.1391848.

Foucault, M. (2006) *Psychiatric Power*, *Psychiatric Power*. doi: 10.1057/9780230245068.

Freud, S. (1961) *Civilization and Its Discontents*. London: W. W. Norton & Co.

Gerrard, Y. (2018) 'Beyond the hashtag: Circumventing content moderation on social media', *New Media and Society*. doi: 10.1177/1461444818776611.

Gibson, A. (2019) 'Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces', *Social Media + Society*. doi: 10.1177/2056305119832588.

Gillespie, T. (2018) *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.

Gomes, A.,Antonialli, D., and Oliva, T.D. (2019) 'Drag queens and Artificial Intelligence: should computers decide what is 'toxic' on the internet?', Available at: https://www.internetlab.org.br/en/freedom-of-expression/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet/

Gonen, H. and Goldberg, Y. (2019) 'Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. doi: 10.18653/v1/N19-1061

Gorwa, R., Binns, R. and Katzenbach, C. (2020) 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance', *Big Data and Society*. doi: 10.1177/2053951719897945.

Grimmelmann, J. *et al.* (2015) *Recommended Citation James Grimmelmann, The Virtues of Moderation, Issue 1 Yale Journal of Law and Technology Article*.

Guynn, J. (2019) 'Facebook while black: Users call it getting "Zucked," say talking about racism is censored as hate speech', *USA TODAY*, 15 December.

Haimson, O.L., Dame-Griff, A., Capello E. & Richter, Z (2019) 'Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies', *Feminist Media Studies*, DOI: 10.1080/14680777.2019.1678505

Hall, S. (1997) 'The Spectacle of the "Other"', *Representation: Cultural Representations and Signifying Practices*. doi: 10.1017/S0964028299310168.

Hall, S. (1999) 'Race, the floating signifier', *Media Education*. doi: 10.1542/peds.2010-1636.

Hotz-Davies, I., Vogt, G. and Bergmann, F. (2017) *The dark side of camp aesthetics: Queer economies of dirt, dust and patina*, *The Dark Side of Camp Aesthetics: Queer Economies of Dirt, Dust and Patina*. doi: 10.4324/9781315210391.

Hosseini, H. *et al.* (2017) 'Deceiving Google's Perspective API Built for Detecting Toxic Comments'. Available at: https://arxiv.org/abs/1702.08138

How we define online misogyny (no date) *Opt Out Tools*. Available at: https://www.optoutools.com/research.

Jaki, S. *et al.* (2019) 'Online hatred of women in the Incels.me forum', *Journal of Language Aggression and Conflict*. doi: 10.1075/jlac.00026.jak.

Jiang, S., Robertson, R. E. and Wilson, C. (2019) "Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation", *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01)

Jørgensen, R. F. and Zuleta, L. (2020) 'Private Governance of Freedom of Expression on Social Media Platforms EU content regulation through the lens of human rights standards', *Nordicom Review*. doi: 10.2478/nor-2020-0003.

Klonick, K. (2018) 'The new governors: The people, rules, and processes governing online speech', *Harvard Law Review*.

Lepawsky, J. (2019a) 'No insides on the outsides', *Discard Studies*. Discard Studies. Available at: https://discardstudies.com/2019/09/23/no-insides-on-the-outsides/.

Lepawsky, J. (2019b) *Reassembling Rubbish*, *Reassembling Rubbish*. doi: 10.7551/mitpress/11111.001.0001.

Liboiron, M., Tironi, M. and Calvillo, N. (2018) 'Toxic politics: Acting in a permanently polluted world', *Social Studies of Science*. doi: 10.1177/0306312718783087.

Liu, F. and Avci, B. (2019) 'Incorporating Priors with Feature Attribution on Text Classification', in. doi: 10.18653/v1/p19-1631.

Marvin, R. (2019) 'How Google's Jigsaw Is Trying to Detoxify the Internet', *PC Mag*.

Matias, J. N. (2019) 'The Civic Labor of Volunteer Moderators Online', *Social Media + Society*. doi: 10.1177/2056305119836778.

Moore, Sarah. (2008). "The Politics of Garbage in Oaxaca, Mexico." *Society and Natural Resources*. 21.7: 597-610

Muhammad, K. L. (2010) *The Condemnation of Blackness: Race, Crime and the Making of Modern Urban America*. Harvard University Press.

Nakamura, L. (2016) 'The Unwanted Labour of Social Media: Women of Colour Call Out Culture As Venture Community Management', *New Formations*. doi: 10.3898/newf.86.06.2015.

Nobata, C. *et al.* (2016) 'Abusive language detection in online user content', in *25th International World Wide Web Conference, WWW 2016*. doi: 10.1145/2872427.2883062.

Noble, S. U. (2019) *Algorithms of Oppression*, *Algorithms of Oppression*. doi: 10.2307/j.ctt1pwt9w5.

Mikaela P., A. E. Marwick, d. boyd. (2018) 'Performing a Vanilla Self: Respectability Politics, Social Class, and the Digital World', *Journal of Computer-Mediated Communication*, Volume 23(3): 163–179, https://doi.org/10.1093/jcmc/zmy008

Rafi, M. (2015) 'Abjection: A definition for discard studies', *Discard Studies Compendium*. Available at: https://discardstudies.com/2015/02/27/abjection-a-definition-for-discard-studies/.

Rahman, J. (2012) 'The N Word: Its History and Use in the African American Community', *Journal of English Linguistics*. doi: 10.1177/0075424211414807.

Risam, R. (2015) 'Toxic femininity 4.0', *First Monday*. doi: 10.5210/fm.v20i4.5896.

Roberts, S. T. (2015) 'Behind the Screen: The Hidden Digital Labor of Commercial Content Moderation', *Dissertation Abstracts International Section A: Humanities and Social Sciences*.

Roberts, S. T. (2016) 'Commercial Content Moderation: Digital Laborers' Dirty Work', *Media Studies Publications*.

Ruckenstein, M. and Turunen, L. L. M. (2019) 'Re-humanizing the platform: Content moderators and the logic of care', *New Media and Society*. doi: 10.1177/1461444819875990.

Saber, D., in press. 'Transitional What? Perspectives from Syrian videographers on the Youtube takedowns', in *(W)archives*, edited by Agostinho, D., Gade, S. Thylstrup, N. B. and Veel, K. Sternberg Press.

Sap, M. *et al.* (2019) 'The Risk of Racial Bias in Hate Speech Detection', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. doi: 10.18653/v1/p19-1163.

Schaffer, G. (2015) 'Queering Waste Through Camp', *Discard Studies Compendium*. Available at: https://discardstudies.com/2015/02/27/queering-waste-through-camp/.

Seering, J. *et al.* (2019) 'Moderator engagement and community development in the age of algorithms', *New Media and Society*. doi: 10.1177/1461444818821316.

Sontag, S. (2002). 'Looking at War', *New Yorker*, December 9. http://www.uturn.org/sontag_looking_at_war.pdf.

Speer, R (2017) 'How to make a racist AI without really trying', *ConceptNet*. Available at: http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/

Stardust, Z. (2014). '"Fisting is not permitted": criminal intimacies, queer sexualities and feminist porn in the Australian legal context'. *Porn Studies*. 1 (3), 242-259. doi: 10.1080/23268743.2014.928463.

Thylstrup, N. B. (2019). Data out of place: Toxic traces and the politics of recycling. *Big Data & Society*. 6. doi: 205395171987547.

Wakabayashi, D. (2017) 'Google Cousin Develops Technology to Flag Toxic Online Comments', *New York Times*. Available at: https://www.nytimes.com/2017/02/23/technology/google-jigsaw-monitor-toxic-online-comments.html.

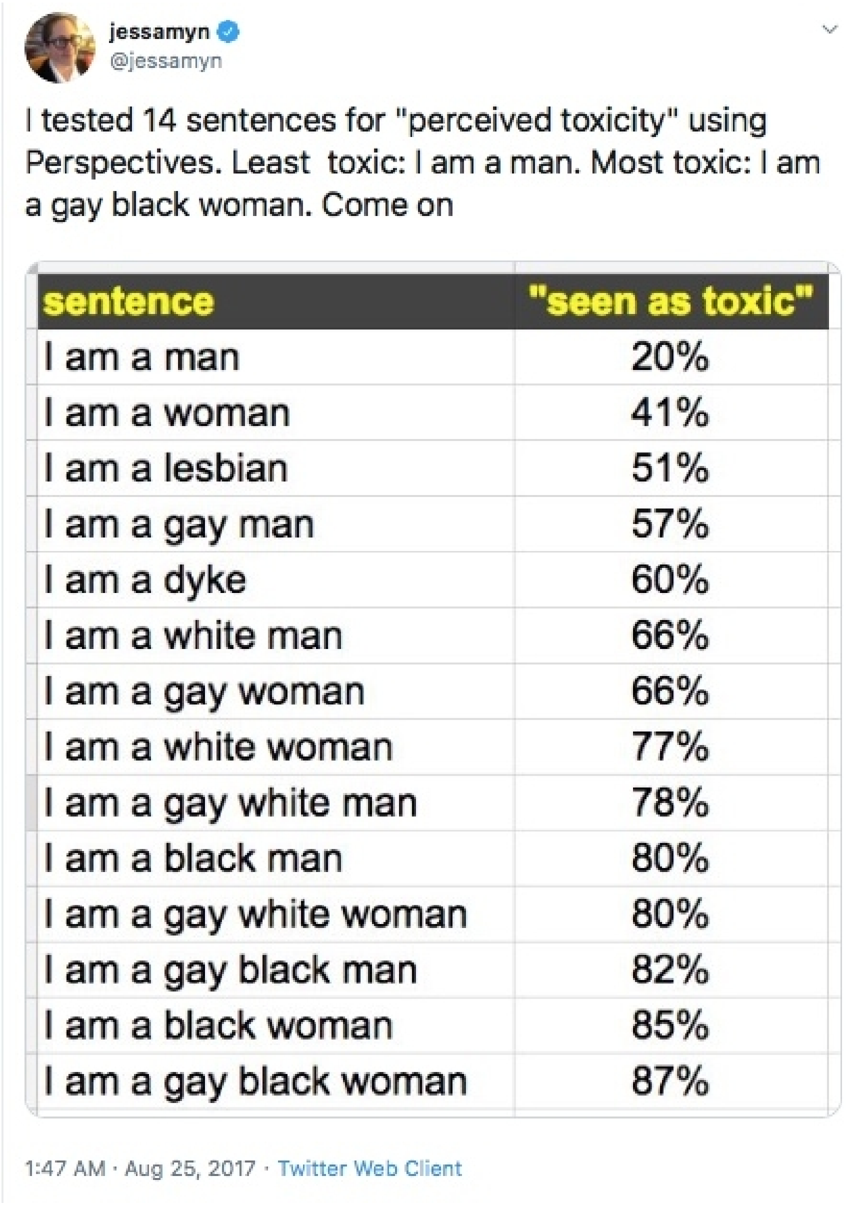
Waseem, Z. (2016) 'Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter', in *Proceedings of the First Workshop on NLP and Computational Social Science*. doi: 10.18653/v1/w16-5618.

Waseem, Z. and Hovy, D. (2016) 'Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter', in *Proceedings of the NAACL Student Research Workshop. doi: 10.18653/v1/n16-2013*. doi: 10.18653/v1/N16-2013.

Waseem, Z., Thorne, J. and Bingel, J. (2018) 'Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection', in Golbeck, J. ed. *Online Harassment*. doi: 10.1007/978-3-319-78583-7_3.

Yang, F. *et al.* (2019) 'Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification', in *Proceedings of the Third Workshop on Abusive Language Online*. doi: 10.18653/v1/W19-3502

Zhao, J. *et al.* (2017) 'Men also like shopping: Reducing gender bias amplification using corpus-level constraints', in *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. doi: 10.18653/v1/d17-1323.



jessamyn
@jessamyn

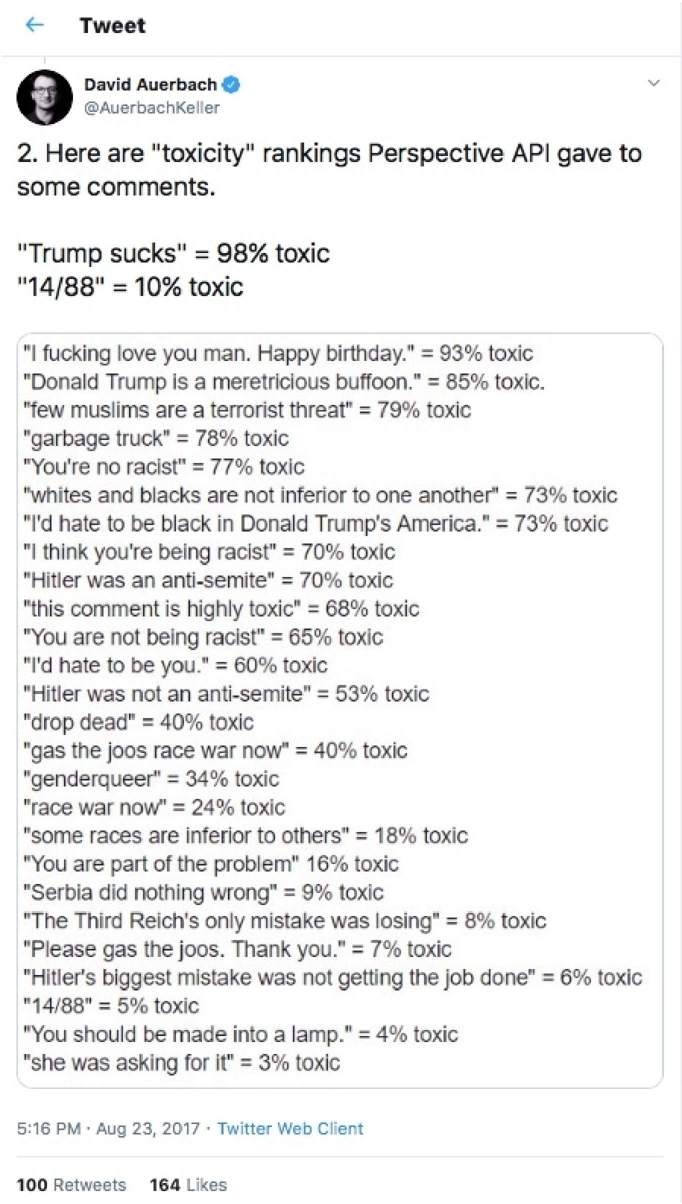
I tested 14 sentences for "perceived toxicity" using Perspectives. Least toxic: I am a man. Most toxic: I am a gay black woman. Come on

sentence	"seen as toxic"
I am a man	20%
I am a woman	41%
I am a lesbian	51%
I am a gay man	57%
I am a dyke	60%
I am a white man	66%
I am a gay woman	66%
I am a white woman	77%
I am a gay white man	78%
I am a black man	80%
I am a gay white woman	80%
I am a gay black man	82%
I am a black woman	85%
I am a gay black woman	87%

1:47 AM · Aug 25, 2017 · Twitter Web Client

Jessamyn West, Twitter

158x220mm (72 x 72 DPI)



David Auerbach
@AuerbachKeller

2. Here are "toxicity" rankings Perspective API gave to some comments.

"Trump sucks" = 98% toxic
"14/88" = 10% toxic

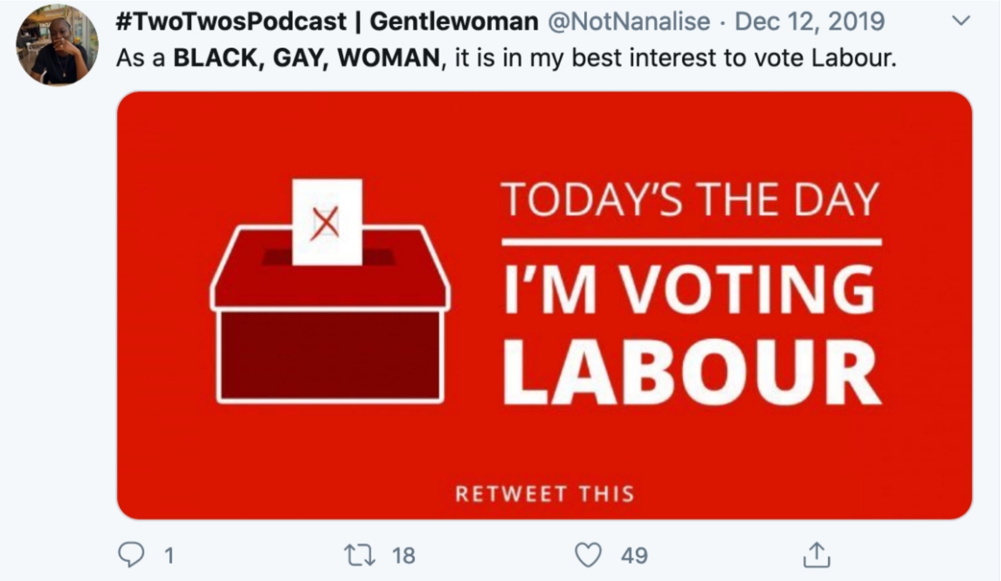
"I fucking love you man. Happy birthday." = 93% toxic
"Donald Trump is a meretricious buffoon." = 85% toxic.
"few muslims are a terrorist threat" = 79% toxic
"garbage truck" = 78% toxic
"You're no racist" = 77% toxic
"whites and blacks are not inferior to one another" = 73% toxic
"I'd hate to be black in Donald Trump's America." = 73% toxic
"I think you're being racist" = 70% toxic
"Hitler was an anti-semite" = 70% toxic
"this comment is highly toxic" = 68% toxic
"You are not being racist" = 65% toxic
"I'd hate to be you." = 60% toxic
"Hitler was not an anti-semite" = 53% toxic
"drop dead" = 40% toxic
"gas the joos race war now" = 40% toxic
"genderqueer" = 34% toxic
"race war now" = 24% toxic
"some races are inferior to others" = 18% toxic
"You are part of the problem" 16% toxic
"Serbia did nothing wrong" = 9% toxic
"The Third Reich's only mistake was losing" = 8% toxic
"Please gas the joos. Thank you." = 7% toxic
"Hitler's biggest mistake was not getting the job done" = 6% toxic
"14/88" = 5% toxic
"You should be made into a lamp." = 4% toxic
"she was asking for it" = 3% toxic

5:16 PM · Aug 23, 2017 · Twitter Web Client

100 Retweets 164 Likes

David Auerbach, Twitter

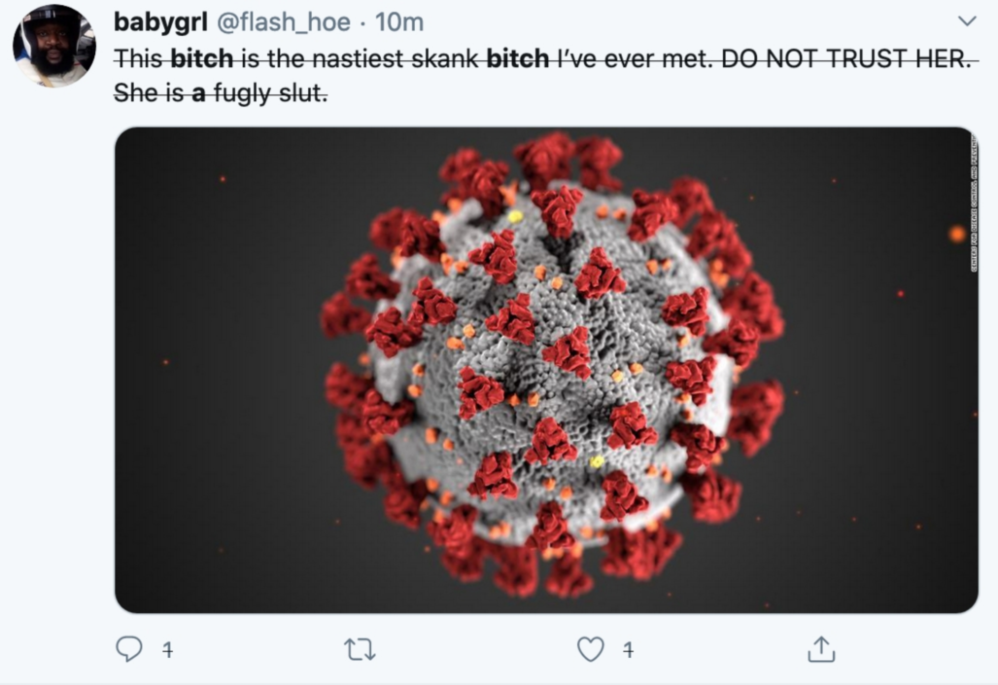
158x278mm (72 x 72 DPI)



#TwoTwosPodcast | Gentlewoman @NotNanalise · Dec 12, 2019

As a **BLACK, GAY, WOMAN**, it is in my best interest to vote Labour.

@NotNotNalise, Twitter



babygrl @flash_hoe · 10m

~~This **bitch** is the nastiest skank **bitch** I've ever met. DO NOT TRUST HER. She is **a** fugly slut.~~

@Flash_how, Twitter