

# False Hopes in Automated Abuse Detection

Tracie Farrell<sup>1</sup>, Soraya Kouadri Mostéfaoui<sup>2</sup>

<sup>1</sup>Knowledge Media Institute (KMi), Open University (OU), Walton Hall, Milton Keynes, MK67AA

<sup>2</sup>School of Computing and Communications (C&C), Open University (OU), Walton Hall, Milton Keynes, MK67AA

## Abstract

The idea of a protected characteristic is supposedly based on the evidence of discrimination against a group of people associated with that characteristic or a combination of those characteristics. However, this determination is political and evolves over time as existing forms of discrimination are recognised and new forms emerge. All the while, these notions are also rooted in colonial practices and legacies of colonialism that *create* and re-create injustice and discrimination against those same “protected” groups. Automated hate-speech detection software is based typically on those political definitions of hate, which are then codified in law. Moreover, the law tends to focus on *classes* of characteristics (e.g. gender, ethnicity), rather than specific characteristics that are particularly targeted by discrimination and hate (being a woman, being Indigenous, Black, Asian, etc.). In this paper, we explore some of the implications of this for hate speech detection, particularly that supported with Artificial Intelligence (AI), and for groups that experience a significant amount of prejudicial hate online.

## Keywords

hate speech detection, protected characteristics, colonialism, artificial intelligence, social justice

## 1. Introduction

The basis for protecting vulnerable groups has shifted across communities and time. In some cases, the need for protection originates in a general concept of dignity for living things and balance within society and nature. In others, religious or political ideologies circumscribe the different groups and things that are “worthy” of our protection. In all cases, legislation is not value neutral. It is designed, written approved and supported, typically, by certain powerful groups of people, which inherently reflects societal biases (political-economy, gender, cultural, racial, etc.). For many modern states, protection for certain vulnerable groups has largely been legislated through anti-discrimination and human rights frameworks, the efficacy of which is debatable in matters of justice.

Most states and regional bodies have attempted to govern the digital world similarly, identifying groups that are vulnerable online and attempting to legislate for their protection. This has involved using computational approaches for identifying, limiting, and in some cases prohibiting certain activities online. Hate-speech is one of the activities that many governments seek to limit, for various reasons and motivations that are not always transparent.

---


*HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 26–27, 2023, Munich, Germany*

✉ [tracie.farrell@open.ac.uk](mailto:tracie.farrell@open.ac.uk) (T. Farrell); [soraya.kouadri@open.ac.uk](mailto:soraya.kouadri@open.ac.uk) (S. K. Mostéfaoui)

ORCID [0000-0002-2386-4333](https://orcid.org/0000-0002-2386-4333) (T. Farrell)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In this paper, we reflect on hate-speech detection algorithms from a socio-political and decolonial perspective, to understand how this particular use of artificial intelligence (AI) and the assumptions on which it is based impact, and in some cases perpetuate, inequality and social justice. Many existing works have covered this subject from the perspective of Fair, Accountable, Transparent and Explainable AI. In this paper, we wish to focus on the origins of hate-speech laws and some of the sociological challenges that have been carried over from the colonial, legislative perspective on hate into the technical implementation of hate-speech detection. In particular, we focus on (mis)perceptions of the “symmetry of equality”, which is apparent in many anti-discrimination frameworks, and how this impacts privilege in online spaces, the suppression of discourse within and between marginalised groups, and the inadequacy of hate-speech identification relative to groups with many intersecting characteristics that lead to discrimination and disadvantage.

## 2. Protected characteristics

A protected characteristic can be described typically as a *definable* characteristic that has been *evidenced* as leading to specific forms of vulnerability or discrimination in society. The law then provides special consideration or provisions pertaining to the *treatment* of individuals holding one or more of these characteristics. While it may appear that governing the treatment of groups with certain characteristics may seem like the most challenging task, defining such characteristics and evidencing them is problematic as well [1]. The UK’s Equality Act and most other anti-discrimination legal frameworks focus on what is called a “grounds-based” approach in which a larger category, such as gender, ethnicity or religion is viewed as the *grounds* upon which a person might be discriminated, rather than highlighting specific groups (e.g. women, people of a minority ethnicity in the country where they live) in need of protection. While this may have made anti-discrimination legislation more pragmatic and palatable to those who might have otherwise rejected it, it may not reflect the lived experience of those most often targeted by discrimination (*ibid.*). The idea of “symmetry” in the grounds that produce inequality creates challenges. There may be groups that are largely not discriminated against, but can use anti-discrimination legal frameworks to receive protection. There will also be groups within society that are vulnerable to discrimination, but not protected specifically under hate speech laws. This includes vulnerable groups, such as asylum seekers or sex workers, whose actions are politicised as undesirable and even threatening to the public, as well as those that are uniquely discriminated in ways that are inscrutable to the hegemonic culture for holding more than one protected characteristic (the concept of intersectionality [2]).

## 3. Origins of hate-speech legislation

An additionally tricky subject is that of how societies have come to legislate against certain types of speech in democracies, particularly in Europe. “Freedom of expression” or “free speech” is a value expressed by many modern democracies as a desirable feature of a democratic society. The decision of how far this extends, *and to whom it extends*, has been hotly debated. In the development of the non-binding Universal Declaration for Human Rights post-World War II, for

example, the allied governments discussed the potential limitations of speech in the interests of national security, preventing intolerance and violence toward certain groups. One argument (championed primarily by then Soviet-aligned states) was that all nations had a duty to prevent any future resurgence of fascism by preventing open, public hate-speech, particularly against different religious or peripheralised ethnic groups. There were concerns, primarily from the US, the UK, that governments could overreach and exploit laws that limit or prohibit certain types of speech in ways that suppress the public [3]. In this instance, the prohibition of hate speech wasn't adopted. However, the binding International Covenant on Civil and Political Rights<sup>1</sup> adopted in 1966 does include an article (Article 20) that addresses the prohibition of “advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence” and “propaganda for war”. This and the International Convention for the Elimination of all Racial Discrimination (which had provisions to criminalise speech that incites ethnicity-based violence) were created within the context of colonial breakdown, apartheid in South Africa and continuing antisemitism in post-war Germany. It may have been easier to make arguments in favour of limiting speech in that context. We certainly have many modern examples where amplification of hateful speech has led to serious physical violence, loss of life and oppression, such as the 1994 genocide in Rwanda [4] and Bosnian War of 1992-1995 [5].

However, to be defined as hate-speech that should be prohibited, many nations have a politically defined set of criteria (such as intent to persecute, posing serious threat to life or inciting crimes against humanity) that set the bar for intervention [6]. This requires both evidence to support one's claim and an *interpretation* of the evidence that will be able to convince anyone passing judgement on whether or not the criteria have been met. That would be challenging, for example, for a globally excluded population to do. Those who have little to no representation in government, even within democratic nations, will not have a role in setting those criteria or evaluating whether or not they have been met. This means that from the start, limiting speech already has the potential to create inequality at the same time as attempting to mitigate it. Indeed, it is crucial to understand how such power dynamics manifest across different groups, and the patterns that emerge.

#### **4. The uselessness of legislating against inequality**

Specific laws and regulations governing the treatment of individuals with certain characteristics can be found in many ancient civilisations. The Code of Hammurabi and the Laws of Eshnunna, both from the ancient Mesopotamia region, for example, included provisions around the treatment of people who were enslaved, women, orphans and abandoned children, and those considered “foreigners” [7]. In many early societies, some of these laws ultimately protected the interests of those who were *subjecting* vulnerable people. Laws that governed the treatment of enslaved people were more likely related to protecting the interests of enslavers, for example. Laws that prevented sexual assault of women most likely protected the interests of husbands and fathers. It is important to recognise this history of protecting one group of people to protect the commercial and personal interests of a different group of people.

In other cases, early laws were an extension of religious values to protect certain vulnerable

---

<sup>1</sup><https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>

groups or, for example, to remain open and hospitable to strangers. The extent to which religious values still influence our concepts of vulnerability and protection is unclear. One can see this in equality laws and frameworks related to the protection of LGBTQAI+ communities, or religious minorities (within the regions they live) [8].

The basis for modern hate-speech detection algorithms are typically legal statutes, such as the Equality Act in the United Kingdom<sup>2</sup> or South Africa<sup>3</sup>, the EU Charter of Fundamental Rights<sup>4</sup>, that relate to certain “protected characteristics”, such as gender, ethnicity, or religious belief. The fact that the internet would be governed by individual nation state’s notions of hate speech is already problematic, but in addition, these initiatives are not developed typically through “bottom-up” engagement with impacted communities or civic participation. Rather, they are derived typically from the moral obligations of single religious and economic traditions or political ideologies, and reinforced by the influence of powerful actors in private and public spheres [9]. As such, they may be viewed as reflecting the existing power asymmetry within the societies in which they emerge, even as they may appear democratic and even universal.

#### 4.1. Limiting hate-speech online

In online spaces, the discourse on regulating speech has continued and the nature of the Internet has presented some challenges to the previously discussed state-defined models of hateful speech [10]. Governments now liaise with companies to ensure that online content viewed by their constituents meets their political and legal standards. This new “technocolonialism” [11] is heavily influencing both public discourse and the web, with limited oversight. The 2016 European Union (EU) Code of Conduct on countering illegal hate speech online, with (then) Facebook, Microsoft, Twitter and YouTube<sup>5</sup> set out a set of commitments expected from social media platforms to address illegal hate speech, defined under the Framework Decision 2008/913/JHA. This decision prohibited the public incitement to “violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin”<sup>6</sup>. Under the Code of Conduct, companies have to demonstrate to the EU how they identify and remove illegal hate-speech in a timely fashion (in some cases 24 hours), their processes for reviewing these practices, and the mechanisms through which platforms can be notified of illegal hate speech. This Code of Conduct is updated periodically, obtaining new commitments from emerging and established social media platforms. There is still tension around regulating hate-speech outside of national or regional territories. The US, for example, has frustrated attempts at multilateral approaches to regulating hate-speech through its commitment to free speech [10]. Hate speech that originates in the US and has an impact elsewhere also goes unregulated.

Ultimately many decisions around what content to moderate and how to intervene is left with companies to decide [12]. Social Media Platforms and Internet Service Providers have their

---

<sup>2</sup><https://www.equalityhumanrights.com/en/equality-act-2010/what-equality-act>

<sup>3</sup><https://www.gov.za/documents/promotion-equality-and-prevention-unfair-discrimination-act>

<sup>4</sup><https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT>

<sup>5</sup>[https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online\\_en](https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combating-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en)

<sup>6</sup>[https://commission.europa.eu/document/download/551c44da-baae-4692-9e7d-52d20c04e0e2\\_n](https://commission.europa.eu/document/download/551c44da-baae-4692-9e7d-52d20c04e0e2_n)

own codes of conduct that they can use to limit or ban certain content [10]. Do companies operating social media platforms have the cultural awareness to spot hate-speech and, if so, is the business model of social media platforms compatible with commitments to limit extreme content? Facebook was determined to have played a key role in the real-world violence against Tamil and Rohingya Muslims, partially due to failures in recognising online threats appropriately and taking decisive action [12]. YouTube's recommendation algorithms have been implicated in many different types of online radicalisation [13], because extreme and shocking content makes user engage [14] and users are recommended through topic detection algorithms to similar content with high engagement.

If it were possible to identify hate-speech accurately and fairly, it might be reasonable to limit it in online spaces that are for public consumption. However, the definition of hate-speech, how serious threats are recognised and evidenced, and the chosen interventions are all deeply contextual issues. Individual nations and private companies do not likely have the will or the capabilities to address this.

## 4.2. Challenges for Automating Hate Speech Detection

To perform automated hate-speech detection, one might search for keywords or phrases, utilise source metadata (like location or demographic information) for additional context, or deploy the use of machine-learning classifiers to spot instances of hate-speech online. Some classifiers make use of deep learning algorithms, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to identify patterns and features in textual data that may indicate hate speech.

There are many challenges to hate speech detection, however, a few key challenges can be underlined to demonstrate how hate speech detection distances us from the originally stated goals of recognising and limiting hate speech. These are: biases inherent in static datasets annotated by humans, the tendency toward focusing on a single characteristic at a time and thus missing intersectional forms of hate, and a general lack of knowledge around the context of hate.

## 4.3. Bias in hate-speech detection datasets

Labelled data is typically necessary to help train and/or evaluate the performance of hate speech detection algorithms. These datasets provide a snapshot in time of how the annotators chosen for the task perceive the data they are labelling. The examples that annotators are given may already introduce bias for certain types or formulations of hateful speech [15, 16]. Studies have shown that annotator experience also shapes how the annotator views instances of hate [17, 18], bringing biases to the labelled dataset. This labelled data goes on to be enriched through other processes and/or subsumed into a pipeline of detection [19, 20, 21], which amplifies existing biases. Language biases, and topic biases [15, 22, 21] can be difficult to fix with technological approaches to debiasing. For example, trying to use models based on English-language hate-speech detection to detect hate in other languages can result in missing "language specific taboo interjections", such as the Spanish use of the word 'puta' (*bitch*) without a misogynistic intention [23]. This occurs with other English-based languages and dialects as well, such as

African American Vernacular English (AAVE) [24].

#### 4.4. The single axis challenge

Hate-speech detection tools also tend to focus on a single-axis (identifying a specific form of hate based on one characteristic, such as race or gender) [25]. This disregards unique forms of hate resulting from intersecting characteristics [2, 25, 26]. Recent studies on automating detection of misogynoir (gendered racism against Black women), for example, indicate that many other contextual features, such as who is talking to whom and in which context, are necessary for understanding and identifying this type of hate [27, 18]. I will discuss context more in the following subsection.

Even deep-learning approaches that can consider multiple forms of hate at the same time mostly look at cumulative single-axis hate. In Founta et al. [28], for example, the authors were still relying heavily on annotated lexicons including explicitly offensive terms related to the single-axis grounds of race, gender, etc. In the example of misogynoir, this approach might capture misogyny, as it is experienced by women, and racism, as it is experienced by Black people, but it will not identify and capture the unique experience of misogynoir, for example, the stereotype of the “angry Black woman”. The resulting *erasure* is the product of a grounds-based approach.

The idea that large language models (LLMs) can resolve these conflicts is still unrealistic. Early experiments using GPT-3 to detect hate around gender and race have shown some limited success [29], however researchers have also identified a bias toward US American values in the outputs and decision making of GPT-3 [30]. Indeed as we have argued in the above sections, because everything related to hate speech tends to be framed around Western values, contexts and concerns, it is no wonder that this also perpetuates in AI (and non-AI) technological tools.

#### 4.5. Lacking Context

The last challenge is context. How do we distinguish public frustration from speech that could meet any of the original, legal definitions of hate speech? One approach is to use patterns in user behaviour to determine intent. For example, in Founta et al. [28], if a user was regularly sending messages that insult other users, they would be categorised as a bully [28]. In Bevendorff et al. [31], focused on identifying the regular spreaders of hate speech, rather than single instances of potentially hateful content.

It may depend, however, on what the user is responding to. In [32], the authors analysed abusive tweets directed at British MPs during the first four months of the COVID pandemic in 2020. They found that preceding some spikes in abuse, there were incidents which would have been worthy of a public outcry, such as improper handling of PPE and miscommunication around lock-downs. That cannot be viewed as hateful speech. However, when an MP who is a Black woman receives repeated personal attacks in response to discussing racism (part of her regular job), that’s something different.

We should note that the MP study above refers to “abuse detection” and not hate-speech detection. There is also “toxicity detection” [16], in which general forms of incivility may be detected alongside what could be called hate-speech. To be able to fully interpret those results,

researchers must attempt to know more about the target and the person targeting that person, to understand if there is a likely instance of hate. Still, similar biases exist in how toxic language is identified and limited [16].

#### **4.6. Implications for targeted groups**

The most obvious impact of hate-speech detection on marginalised groups is over-surveillance, which results in limiting their own activity online. Research indicates, for example, that Black men are more likely to have their content labelled as toxic or abusive because of different language conventions [26, 24]. Black women have also been shown to experience disproportionate influence from content moderation, particularly when discussing racism [33]. Transgender users have been blocked for adult content when discussing their personal experiences [33].

Yes, users learn how to evade detection algorithms [34], for example, through use of euphemistic speech [35] or neologisms [20]. However, researchers come up with new ideas to spot evasion and the cycle continues.

#### **4.7. Other impacts on marginalised people**

It should be noted that content moderation approaches, even when partially automated, still include the use of human labour to screen and remove content. This work, which has been described as traumatising and poorly compensated [36], is often outsourced to labourers in the Philippines and India where the moderators receive no mental health provisions or support [37]. Once again, the legacy of colonialism looms large in the ways that undesirable labour is outsourced to poorer communities with less global influence, so that those with relative influence and power can see a more comforting and sanitised version of the web.

### **5. Conclusion**

In this paper, we have described how the legal, political definition of hate-speech and the protection of vulnerable populations through a grounds-based approach can lead to inequalities through (mis)perceptions of “symmetry” without attention to power gifted by the legacy of colonialism and capitalism. We explained how these inequalities appear in the online sphere, through non-context aware hate and toxicity detection that limits the activities of marginalised and peripheralised groups. Finally, we highlighted some of the unique challenges that arise when large, powerful tech companies are in charge of how hate-speech is defined and what speech is worthy of censure. While there is certainly a case for limiting hateful speech on social media platforms, the currently deployed methods are not particularly ethical or fair in preventing inequality for people who are marginalised. We need more transformative, empowered and community-based solutions that involve dialogue with those truly impacted by hateful speech online. In particular, we need to focus on impacts which translate into or are derived from power asymmetry offline. A continued paternalistic, and technocolonialist approach to limiting speech online will at best resolve the whole into a box-ticking exercise and at worst exacerbate inequality.

## Acknowledgments

This work was funded by a UKRI Future Leaders Fellowship (Round Six) MR/W011336/1.

## References

- [1] K. Malleson, Equality law and the protected characteristics, *The Modern Law Review* 81 (2018) 598–621.
- [2] K. W. Crenshaw, *On intersectionality: Essential writings*, The New Press, 2017.
- [3] J. Mchangama, The sordid origin of hate-speech laws, *Policy Review* (2011) 45.
- [4] C. L. Kellow, H. L. Steeves, The role of radio in the rwandan genocide, *Journal of communication* 48 (1998) 107–128.
- [5] J. Pálmadóttir, I. Kalenikova, Hate speech an overview and recommendations for combating it, *Icelandic Human Rights Centre* (2018) 1–27.
- [6] G. S. Gordon, Hate speech and persecution: A contextual approach, *Vand. J. Transnat'l L.* 46 (2013) 303.
- [7] R. H. Hiers, *A Nation of Immigrants: Sojourners in Biblical Israel's Tradition and Law*, Wipf and Stock Publishers, 2021.
- [8] R. Moon, *Putting faith in hate: when religion is the source or target of hate speech*, Cambridge University Press, 2018.
- [9] P. Carls, Free speech and the protection of human dignity in canada, germany, and the united states: The moral legitimation of competing discourses, *Canada* 150 (2020) 203–220.
- [10] J. Banks, Regulating hate speech online, *International Review of Law, Computers & Technology* 24 (2010) 233–239.
- [11] M. Madianou, Technological futures as colonial debris: 'tech-for-good' as technocolonialism (2022).
- [12] Z. Laub, Hate speech on social media: Global comparisons, *Council on foreign relations* 7 (2019).
- [13] K. Roose, The making of a youtube radical, *The New York Times* 8 (2019).
- [14] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, *science* 359 (2018) 1146–1151.
- [15] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, *PloS one* 14 (2019) e0221152.
- [16] T. Garg, S. Masud, T. Suresh, T. Chakraborty, Handling bias in toxic speech detection: A survey, *ACM Computing Surveys* (2022).
- [17] Z. Waseem, 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*, 2016, pp. 138–142.
- [18] J. Kwarteng, S. C. Perfumi, T. Farrell, A. Third, M. Fernandez, Misogynoir: challenges in detecting intersectional hate, *Social Network Analysis and Mining* 12 (2022) 166.
- [19] G. Mou, P. Ye, K. Lee, Swe2: Subword enriched and significant word emphasized framework for hate speech detection, in: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1145–1154.



- [20] T. Farrell, O. Araque, M. Fernandez, H. Alani, On the use of jargon and word embeddings to explore subculture within the reddit’s mansphere, in: 12th ACM Conference on web science, 2020, pp. 221–230.
- [21] E. W. Pamungkas, V. Basile, V. Patti, A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection, *Information Processing & Management* 58 (2021) 102544.
- [22] F. Alkomah, X. Ma, A literature review of textual hate speech detection methods and datasets, *Information* 13 (2022) 273.
- [23] D. Nozza, Exposing the limits of zero-shot cross-lingual hate speech detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 907–914.
- [24] P. Reyer Lobo, Bias in hate speech and toxicity detection, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022, pp. 910–910.
- [25] A. L. Hoffmann, Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse, *Information, Communication & Society* 22 (2019) 900–915.
- [26] J. Y. Kim, C. Ortiz, S. Nam, S. Santiago, V. Datta, Intersectional bias in hate speech and abusive language datasets, arXiv preprint arXiv:2005.05921 (2020).
- [27] J. Kwarteng, S. C. Perfumi, T. Farrell, M. Fernandez, Misogynoir: public online response towards self-reported misogynoir, in: Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining, 2021, pp. 228–235.
- [28] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, I. Leontiadis, A unified deep learning architecture for abuse detection, in: Proceedings of the 10th ACM conference on web science, 2019, pp. 105–114.
- [29] K.-L. Chiu, A. Collins, R. Alexander, Detecting hate speech with gpt-3, arXiv preprint arXiv:2103.12407 (2021).
- [30] R. L. Johnson, G. Pistilli, N. Menéndez-González, L. D. D. Duran, E. Panai, J. Kalpokiene, D. J. Bertulfo, The ghost in the machine has an american accent: value conflict in gpt-3, arXiv preprint arXiv:2203.07785 (2022).
- [31] J. Bevendorff, B. Chulvi, G. L. De La Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, et al., Overview of pan 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, Springer, 2021, pp. 419–431.
- [32] T. Farrell, G. Gorrell, K. Bontcheva, Vindication, virtue, and vitriol: A study of online engagement and abuse toward british mps during the covid-19 pandemic, *Journal of Computational Social Science* 3 (2020) 401–443.
- [33] O. L. Haimson, D. Delmonaco, P. Nie, A. Wegner, Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–35.
- [34] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, N. Asokan, All you need is” love” evading hate speech detection, in: Proceedings of the 11th ACM workshop on artificial intelligence and

security, 2018, pp. 2–12.

- [35] R. Magu, J. Luo, Determining code words in euphemistic hate speech using word embedding networks, in: Proceedings of the 2nd workshop on abusive language online (ALW2), 2018, pp. 93–100.
- [36] S. T. Roberts, Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste, *Wi: journal of mobile media* (2016).
- [37] S. W. Baek, Ignored and deleted: Understanding content moderators as racialized media of social network services (2022).