Michigan Technology Law Review

Article 4

2023

Coca-Cola Curses: Hate Speech in a Post-Colonial Context

Brittan Heller
Affiliate, Stanford Cyber Policy Center and Senior Fellow, Atlantic Council

COCA-COLA CURSES: HATE SPEECH IN A POST-COLONIAL CONTEXT

Brittan Heller[*]

ABSTRACT

Hate speech is a contextual phenomenon. What offends or inflames in one context may differ from what incites violence in a different time, place, and cultural landscape. Theories of hate speech, especially Susan Benesch's concept of "dangerous speech" (hateful speech that incites violence), have focused on the factors that cut across these paradigms. However, the existing scholarship is narrowly focused on situations of mass violence or societal unrest in America or Europe.

This paper discusses how online hate speech may operate differently in a postcolonial context.[1] *While hate speech impacts all societies, the global South—Africa in particular—has been sorely understudied. I posit that in postcolonial circumstances, the interaction of multiple cultural contexts and social meanings form concurrent layers of interpretation that are often inaccessible to outsiders. This study expands the concept of online harms by examining the political, social, and cultural dimensions of data-intensive technologies.*

The paper's theories are informed by fieldwork that local partners and I conducted in Kasese, Uganda in 2019–2020, focusing on social unrest and lethal violence in the region following the 2016 elections. The research, completed with assistance from the Berkeley Human Rights Clinic, included examining the background and circumstances of the conflict; investigating social media's role in the conflict; designing a curriculum around hate speech and disinformation for Ugandan audiences; creating a community-sourced lexicon of hateful terms; and incorporating community-based feedback on proposed strategies for mitigating hate speech and disinformation.

---

* Affiliate, Stanford Cyber Policy Center and Senior Fellow, Atlantic Council. I would like to thank Alexa Koenig, Andrea Lampros, Musoki Elizabeth, Johncation Muhindo, Larry Diamond, Nicholas Opiyo, Allen Weiner, Mark Lemley, Nathaniel Gleicher, Danielle Citron, Nathan Matias, Sonnet Phelps, Susan Benesch, the residents of Kasese, the Mandela-Washington Scholars Program, the U.S. Embassy in Kampala, and the Privacy Law Scholars Conference.

1. "Postcolonial," as used in this paper, refers to a theoretical approach in various disciplines that is concerned with the lasting impact of colonization in former colonies. *See, e.g.*, ANNETTE KUHN & GUY WESTWELL, A DICTIONARY OF FILM STUDIES (Oxford Univ. Press 1st ed. 2012), https://www.oxfordreference.com/display/10.1093/acref/9780199587261.001.0001/acref-9780199587261-e-0543.

I begin this with a literature review of legal theory around hate speech, with a particular focus on Africa, and then turn to the legal context around hate speech and social media use in Uganda, examining how the social media landscape fueled past conflicts. Then I explain my Kasese fieldwork and the study's methodology, before describing initial results. I follow with a discussion of applications to industry, specifically how hate speech is defined and treated by Meta's Facebook, the dominant social media provider in Kasese. It progresses to a discussion of the implications of the study results and legal and policy recommendations for technology companies stemming from these findings.

Importantly, I apply the research findings to expand existing scholarship by proposing a new sixth "hallmark of dangerous speech" to augment Benesch's paradigm. Adding "calls for geographic exclusion" as a new qualifier for dangerous speech stems from the particular characteristics embodied by postcolonial hate speech. Examples from the Kasese study illustrate how this phenomenon upends platforms' expectations of hate speech—which may not consider "Coca-Cola bottle" to be an epithet. The application of this new hallmark will create a more inclusive understanding of hate speech in localized contexts.

This paper's conclusions and questions may challenge platforms that must address hate speech and content moderation at a global scope and scale. It will examine the prevalence and role of social media platforms in Africa, and how these platforms have provided resources and engagement with civil society in these regions.

TABLE OF CONTENTS

I. INTRODUCTION

Recalling my fieldwork to create a lexicon of hate speech, in Kasese, Uganda,[2] one moment stands out above the rest. After three days of training local religious, tribal, and civic leaders on what hate speech means and how it functions, a hesitant participant stood up to provide a local example: a Kasese politician had been referred to as a "Coca-Cola bottle." Suddenly the room burst into noise and chaos, with loud tones of outrage and nervous laughter.

This was surprising. Initially, the example did not cleanly fit prevailing scholarship on hate speech. It was unclear how "Coca-Cola bottle" targeted a person for the immutable characteristics that hate speech typically focuses on, like race, gender, or ethnicity.

However, after explanation from local Ugandan partners and further investigation, the meaning of the epithet became clear. The candidate was from the Bakonzo people, descended from a mountainous tribe, which some in the crowd referred to as "pygmy" to help me try to understand. This candidate's tribe was historically shorter than his opponent's, just as a bottle is much

2. *See infra* Appendix I.

smaller than a man. Similarly, Coca-Cola's dark brown color emphasized the candidate's skin, darker than his opponent's.

On its face, this remark might be considered superficial mockery of the candidate's physical characteristics. But the insult went deeper. "Coca-Cola bottle" alluded to the film *The Gods Must Be Crazy*.[3] The plot's conceit is that an African tribesman, finding a glass Coca-Cola bottle in the desert, becomes convinced it is a gift from the gods. As the glass bottle provokes jealousy and unrest in his tribe, the man decides it must be destroyed. The movie has drawn criticism for stereotypical depictions of "Bushmen" and perpetuating colonial attitudes about the ignorance of tribal peoples. In Kasese, calling a political opponent a Coca-Cola bottle disparaged him as ignorant, a colonial tool, or possessing characteristics of a stereotype—or perhaps all these attributes at once.

Finally, calling someone a Coca-Cola bottle fits with the theory of dangerous speech. Dehumanizing and often hateful speech is used to normalize potential violence against a group. This type of rhetoric follows patterns spanning location, culture, and time. One such trope casts a group and its members as less than human through comparisons to vermin, disease, or garbage—just like a Coca-Cola bottle is thrown away as trash once the drink has been consumed.

"Coca Cola bottle" was no anomaly. Another striking example which came up repeatedly in focus groups was the identification of "white beauty standards" as hate speech. This was not a term, but a social milieu—and yet, during several discussions, Ugandan civic leaders unanimously agreed it constituted hate speech. Studying this pattern, the problem became apparent: it was not a misunderstanding by Kasese residents, but rather limitations in existing conceptions of what hate speech looks like, and how it functions. In a postcolonial environment, what is hateful may manifest differently, and be more inaccessible to outsiders.

Why should we focus on hate speech in postcolonial contexts? Aside from the cost in human suffering from hate speech and derogation, many of the most brutal ethnic conflicts in recent memory have arisen in postcolonial environments and were built off social legacies of colonial rule. From the radio in Rwanda and Sudan to social media in Myanmar, Kenya, Sri Lanka, and India, these communications platforms spread hateful narratives to fuel conflicts.[4]

This paper posits that in a postcolonial context, analyzing hate speech warrants a unique approach. Culture and context imposed under colonization may exist in fundamental tension with, or in opposition to, the history, tradition, and culture of the colonized. In this way, a Coca-Cola bottle can be seen

3. *See generally* THE GODS MUST BE CRAZY (C.A.T. Films 1980).

4. *See infra* Section I.A. (Examples include Rwanda, Cambodia, Sudan, Myanmar, and present-day violence occurring in India.)

as both hateful and not hateful. The duality sublimates postcolonial hate speech in subtle but volatile ways; in other words, observers from outside the colonized culture may not realize the full impact of hate speech, especially if it can be mitigated by the more innocuous understanding of the term in the colonizers' context, until the effects of such speech catalyze ethnic, religious, or political violence.[5]

Because hate speech is a contextual animal, there is a fundamental challenge for content moderation on social media platforms. This example from Kasese illustrates how the same content can both elicit local audiences' outrage and dodge hate speech detection by major social media companies. It is doubtful that calling someone a Coca-Cola bottle would violate the terms of service of a social media company utilizing a predominantly American perspective, unless the reference was seen as an infringement of intellectual property. These layers of social meaning likely would have evaded automated content moderation filters.

A rough reaction to this problem might be to set automatic filters to capture "Coca-Cola bottle" as a hateful symbol. Even with geofencing, this approach risks over-enforcement.[6] Most hate speech on Facebook is caught by artificial intelligence (AI) before it hits user feeds,[7] but it would be challenging, based on AI's general inability to interpret context, to understand when "Coca-Cola bottle" is used in a hateful manner rather than to refer to the literal beverage.[8] Even if the platforms could achieve a proper balance, the Coca-Cola Company may push back because of the potential negative impact on its brand through association with a slur.

Similarly, the comparison is not likely to alert human moderators to flag the post as hate speech, especially if they are unfamiliar with local mores. In a postcolonial society like Kasese, Uganda, where concurrent levels of social meaning exist in tension with one another, the full implications of the context are veiled.

Addressing this form of hate speech cannot counter the colonial history that has led to countless modern-day conflicts, but we can understand what it means, how it manifests, and how it helps fuel hateful narratives and ethnic tensions. Understanding how online hate speech uniquely functions in

5. Susan Benesch & Jonathan Leader Maynard, *Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention*, 9 GENOCIDE STUD. & PREVENTION 69, 80–81 (2016), (an example of this includes the use of "cockroach" during the Rwandan genocide).

6. *See, e.g.*, Emma Llanso, *Human Rights NGOs in Coalition Letter to GIFCT*, CTR. FOR DEMOCRACY & TECH.: FREE EXERCISE (July 30, 2020), https://cdt.org/insights/human-rights-ngos-in-coalition-letter-to-gifct.

7. Arcadiy Kantor, *Measuring Our Progress Combating Hate Speech*, META: NEWSROOM (Nov. 19, 2020) (Discussing Facebook), https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech.

8. *See* Brittan Heller, Opinion, *Is This Frog a Hate Symbol or Not?* N.Y. TIMES (Dec. 24, 2019), https://www.nytimes.com/2019/12/24/opinion/pepe-frog-hate-speech.html.

environments like Kasese may help us build more effective tools to combat it. This can also help us become more sensitive to the ripples that colonialism continues to create in societies, long after the colonizers leave, and that evolving technological mediums refract anew.

A. *Human Rights and Hate Speech*

Divisive content has long been promoted using state-of-the-art technology, and the past few years have been no exception.[9] The growth of social media has spotlighted online hate speech as a heated subject of concern for the international community.[10] Hate speech distributed via social media has increasingly been linked to instances ranging from individual acts of violence to genocide.[11] For example, in March 2018, United Nations investigators declared that Facebook had played a "determining role" in the ethnic cleansing of Myanmar's minority Rohingya Muslims by the country's military and allied Buddhist groups.[12] As many as 10,000 Rohingya were killed, and more than 650,000 fled as refugees to neighboring Bangladesh.[13] Facebook commissioned a human rights impact assessment to determine the company's impact on ethnic conflict in Myanmar; the findings linked the organic content (Facebook's term for user-generated content) and content moderation policies and capabilities of the platform to the incitement of violence.[14]

Other violent incidents demonstrated to the world how online hate speech could result in increased discrimination, strife, and the targeting of minorities. Facebook is the world's largest social media company, with a highly intricate platform architecture, an evolving policy apparatus, and an operational

---

9. *See, e.g.*, HEIDI J. S. TWOREK, NEWS FROM GERMANY 141–169 (Harvard Univ. Press 2019).

10. The United Nations launched a Strategy and Plan of Action on Hate Speech in 2020, recognizing "that over the past 75 years, hate speech has been a precursor to atrocity crimes, including genocide, from Rwanda to Bosnia to Cambodia." *United Nations Strategy and Plan of Action on Hate Speech*, UNITED NATIONS OFFICE ON GENOCIDE PREVENTION AND THE RESPONSIBILITY TO PROTECT (May 2019), https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf.

11. Alex Warofka, *An Independent Assessment of the Human Rights Impact of Facebook in Myanmar*, FACEBOOK (Aug. 26, 2020, 12:30 AM), https://about.fb.com/news/2018/11/myanmar-hria. *See also Myanmar: Facebook's Systems Promoted Violence Against Rohingya; Meta Owes Reparations*, AMNESTY INT'L (Sept. 29, 2022), https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/.

12. Tom Miles, *U.N. Investigators Cite Facebook Role in Myanmar Crisis*, REUTERS (Mar. 12, 2018), https://www.reuters.com/article/us-myanmar-rohingya-facebook-idUSKCN1GO2PN.

13. *Id*. *See also* Catesby Holmes, *Myanmar Charged with Genocide of Rohingya Muslims: 5 Essential Reads*, THE CONVERSATION (Dec. 11, 2019), https://theconversation.com/myanmar-charged-with-genocide-of-rohingya-muslims-5-essential-reads-128742.

14. Dunstan Allison-Hope, *Our Human Rights Impact Assessment of Facebook in Myanmar*, BSR (Nov. 5, 2018), https://www.bsr.org/en/our-insights/blog-view/facebook-in-myanmar-human-rights-impact-assessment.

ecosystem deployed at massive scale.[15] Beginning with its May 2018 reporting on Myanmar, the company contracted with outside auditors to complete assessments of its impact on human rights in other countries with a volatile social media climate arising from online behavior on Facebook, Messenger, WhatsApp, and Instagram.[16] These audits included: Sri Lanka (issued November 2018),[17] Indonesia (issued December 2018),[18] and Cambodia (issued December 2019).[19] In addition, Facebook issued a Civil Rights audit, focused on platforms in the United States, in July 2020.[20] Future assessments are reported to focus on Asia.[21]

Despite consensus on the grave challenges of hate speech, there is no legal, academic, or colloquial agreement on its definition.[22] It would be difficult to define the concept in a way that is sufficiently broad and flexible to capture the varied and evolving forms of hate speech while also remaining narrowly tailored enough not to resemble censorship or otherwise impinge on freedom of expression. Academics and legal experts have heavily debated an appropriate balance.[23]

---

15. *See* S. Dixon, *Number of Monthly Active Facebook Users Worldwide as of 3rd Quarter 2022*, STATISTA (Oct. 27, 2022), https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide (Facebook had 2.79 billion monthly users as of 2020.).

16. *See* Allison-Hope, *supra* note 14.

17. *Assessing the Human Rights Impact of Facebook's Platforms in Sri Lanka*, ARTICLE ONE (2018), https://about.fb.com/wp-content/uploads/2020/05/Sri-Lanka-HRIA-Executive-Summary-v82.pdf.

18. *Assessing the Human Rights Impact of Facebook's Platforms in Indonesia*, ARTICLE ONE (2018), https://articleoneadvisors.com/wp-content/uploads/2023/01/Indonesia-HRIA_-Executive-Summary_FINAL.pdf.

19. *Human Rights Impact Assessment: Facebook in Cambodia*, BSR (Dec. 2019), https://about.fb.com/wp-content/uploads/2020/05/BSR-Facebook-Cambodia-HRIA_Executive-Summary2.pdf.

20. Laura W. Murphy & Megan Cacace, *Facebook's Civil Rights Audit - Final Report*, FACEBOOK (July 8, 2020), https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf.

21. Miranda Sissors & Alex Warofka, *An Update on Facebook's Human Rights Work in Asia and Around the World*, FACEBOOK (May 12, 2020), https://about.fb.com/news/2020/05/human-rights-work-in-asia.

22. *See* Combating Racist Hate Speech (CERD Recommendation No. 35), ¶ 46, U.N. Doc. CERD/C/GC/35, (Sept. 26, 2013); European Commission against Racism and Intolerance, *ECRI General Policy Recommendation No. 15 on Combating Hate Speech*, COUNCIL OF EUROPE, 16 (Dec. 8, 2015). *See also* U.N. Secretary-General's Remarks at the Launch of the United Nations Strategy and Plan of Action on Hate Speech (June 18, 2019), https://www.un.org/sg/en/content/sg/statement/2019-06-18/secretary-generals-remarks-the-launch-of-the-united-nations-strategy-and-plan-of-action-hate-speech-delivered; Opening Statement by U.N. High Commissioner for Human Rights, 41st Session of the Human Rights Council (June 24, 2019), https://www.ohchr.org/en/statements/2019/06/41st-session-human-rights-council.

23. *See, e.g.*, David Kaye, author of the *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UNITED NATIONS HUM. RTS. COUNCIL 10 (Apr. 6, 2018), https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement.

B. *Hate Speech under International Law*

International law provides guidance under foundational declarations, treaties, and advisory documents, about permissible and impermissible kinds of speech under the umbrella of freedom of expression, and sheds some light on legal consensus on integrating freedom of expression with safety.[24] This was first enshrined in Article 19 of the Universal Declaration of Human Rights ("UDHR"), which states: "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."[25]

Similarly, Article 19 of the International Convention for Civil and Political Rights ("ICCPR") states: "(1) Everyone shall have the right to hold opinions without interference. (2) Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice."[26] Article 20 of the ICCPR provides for certain restrictions on free speech. It states: "(1) Any propaganda for war shall be prohibited by law; (2) Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."[27]

Additionally, the International Convention on the Elimination of all Forms of Racial Discrimination ("CERD") addresses prohibitions on hateful speech in Article 4(a), requiring governments to outlaw "all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin."[28]

Overall, rights may be limited only on the basis of specific conditions prescribed in an applicable treaty. To be justified, any limitation of the right to freedom of expression must meet the CERD three-part test[29] requiring that: (i) the limitation must be provided for in law;[30] (ii) it must pursue a legitimate

24. Additionally, other fundamental human rights implicated by online speech may include freedom of assembly, the right to an education, the right to an adequate standard of living, the right to privacy, and the right to health and welfare.

25. G.A. Res. 217 (III) A, Universal Declaration of Human Rights, § 19 (Dec. 10, 1948).

26. International Covenant on Civil and Political Rights § 19, *adopted* Dec. 19, 1966, 999 U.N.T.S. 171 [hereinafter *ICCPR*].

27. *Id.* at § 20.

28. International Convention on the Elimination of All Forms of Racial Discrimination art. 4(a), *opened for signature* Mar. 7, 1966, 660 U.N.T.S. 195.

29. AVANI SINGH, LEGAL STANDARDS ON FREEDOM OF EXPRESSION: TOOLKIT FOR THE JUDICIARY IN AFRICA 50 (United Nations Educational, Scientific, and Cultural Organization 2018).

30. Freedoms of Opinion and Expression (CERD Recommendation No. 35), ¶ 24, U.N. Doc. CERD/C/GC/34, (Sept. 12, 2011).

aim; [31] and (iii) it must be necessary for a legitimate purpose.[32] While Article 20 does not define hate speech, under this rubric, hate speech falls outside of the realm of protected speech. The sentiment is echoed in other seminal bodies. Article 19(3) of the ICCPR[33] contains restrictions on the right to freedom of expression in both treaties. Article 19(3) of the ICCPR states that: "The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (*ordre public*), or of public health or morals."[34] These limitations are echoed in Article 9(2) of the African Charter on Human and Peoples' Rights[35] ("African Charter").

Even though major human rights treaties and regional treaty bodies for Africa were drafted before the internet, the rights contained within them still apply to online spaces. Article 19 of the ICCPR asserts that the right to freedom of expression is applicable to any media and regardless of frontiers.[36] Furthermore, the African Charter[37] asserts that individuals' rights to freedom of expression and freedom of information must be protected.[38]

As hate speech often targets minority populations, examining African sources in international law may demonstrate the positive obligation on States to protect all voices. Principle III of the Declaration on Principles on Freedom of Expression in Africa details the close relationship between diversity and free speech:

> Freedom of expression imposes an obligation on the authorities to take positive measures to promote diversity, which include among other things:
>
> - Availability and promotion of a range of information and ideas to the public;
>
> - Pluralistic access to the media and other means of communication, including by vulnerable or marginalized groups, such as women, children and refugees, as well as linguistic and cultural groups;

31. SINGH, *supra* note 29, at 50.
32. *Id*.
33. *ICCPR*, *supra* note 26, at art. 19(3).
34. *Id.*
35. African Charter on Human and Peoples' Rights art. 9(2), *adopted* June 1, 1981, 1520 U.N.T.S. 217 (entered into force Oct. 21, 1986) [hereinafter *African Charter*].
36. *ICCPR, supra* note 26, at art. 19.
37. *African Charter*, *supra* note 35.
38. *Id.* at art. 9–10.

- The promotion and protection of African voices, including through media in local languages; and

- The promotion of the use of local languages in public affairs, including in the courts.[39]

Along with positive rights, there are also negative rights under international law, if specific conditions warranting prohibitions are met. Any restriction or penalty on speech labelled as 'hate speech' must still conform to the three-part test, referenced previously, for a lawful limitation or restriction of the right to freedom of expression.[40] Holdings by African regional and sub-regional international courts have applied the three-part test to determine if limitations on the right to freedom of expression were warranted under law. For example, in *Zongo v. Burkina Faso*, the African Court held that the state had violated the right to freedom of expression under Article 9 of the African Charter by failing to investigate and prosecute the murderers of Zongo, a journalist.[41] In *Konaté v. Burkina Faso*, the African Court held Article 9's guarantee of freedom of expression was infringed by aspects of the criminal defamation law, particularly provisions that imposed imprisonment as potential punishment.[42] Finally, in *Federation of African Journalists and Others v. The Gambia,* the ECOWAS[43] Court of Justice ordered the Gambia to immediately repeal or amend its laws on criminal defamation, sedition, and false news because these statutes did not comply with the State's obligation to protect free speech under international law.[44]

Other sources of international jurisprudence, such as international criminal tribunals that charge genocide, do not clarify issues surrounding the definition of hate speech, even when incitement to violence or the targeting of minorities is at stake.[45] More is left undefined than is defined.[46] One brief and rare definition of hate speech comes from the holdings of the International Criminal Tribunal for Rwanda, under the *Nahimana et al. Appeal Judgment*, which distinguishes between direct and public incitement to commit genocide

39. Afr. Comm'n on Human & Peoples' Rights Res. 62 (XXXII)02, §III (Oct. 23, 2002).

40. SINGH, *supra* note 29, at 50.

41. Oliver Windridge, *Introductory Note to Zongo v. Burkina Faso, Judgment & Judgment on Reparations (Afr. Ct. H.P.R.)*, 56 INT'L LEGAL MATERIALS 1091, 1092 (2017).

42. Dinah Shelton, *Konaté v. Burkina Faso*, 109 AM. J. OF INT'L L. 630, 632 (2015).

43. This is the acronym for the Economic Community of West African States.

44. Federation of African Journalists v. Gambia, ECW/CCJ/JUD/04/18 61–62 (ECOWAS Ct. of Just. 2018), http://www.courtecowas.org/wp-content/uploads/2019/02/ECW_CCJ_JUD_04_18.pdf.

45. *See generally* RICHARD ASHBY WILSON & MATTHEW GILLET, THE HARTFORD GUIDELINES ON SPEECH CRIMES IN INTERNATIONAL CRIMINAL LAW 10, 13–17, 24–30 (Peace and Just. Initiative 2018).

46. *See generally id.*

and "hate speech in general (or inciting discrimination or violence)."[47] No further explanation is given.[48] In two other Bosnian cases, the *Šešelj Trial Judgment*[49] and the *Šešelj Appeal Judgment*[50], the International Residual Mechanism for the International Criminal Tribunal did not set out a clear definition of hate speech.

While the provisions in the ICCPR reference "hatred," they do not use the terminology "hate speech."[51] Still, under Article 20, three types of speech can be distinguished, which may also apply to hateful speech: "Hate speech that must be prohibited by States (article 20(2) of the ICCPR); hate speech that may be prohibited by States (such as article 19(3) of the ICCPR); and lawful hate speech that should be protected from restriction, but nevertheless raises concerns in terms of intolerance and discrimination, and may merit a critical response by the State (such as article 19(2) of the ICCPR)."[52]

States bear primary responsibility for protecting their citizens' human rights, and accordingly, domestic law is privileged over other forms of international law like treaties, conventions, or customs in practice.[53] In the African context, domestic law is also given primacy over international law. However, the African Commission on Human and Peoples' Rights ("ACHPR"), established by the African Charter to promote and protect the rights enshrined therein, takes a wider view of the range of sources when determining a matter before it. Under the African Charter, Article 60:

---

47. Prosecutor v. Nahimana, Case No. ICTR-99-52-A, Appeal Judgement, § 692 (Nov. 28, 2007), https://ucr.irmct.org/LegalRef/CMSDocStore/Public/English/Judgement/NotIndexable/ICTR-99-52/MSC31299R0000555179.PDF ("The Appeals Chamber considers that there is a difference between hate speech in general (or inciting discrimination or violence) and direct and public incitement to commit genocide. Direct incitement to commit genocide assumes that the speech is a direct appeal to commit an act referred to in Article 2(2) of the Statute; it has to be more than a mere vague or indirect suggestion. ln most cases, direct and public incitement to commit genocide can be preceded or accompanied by hate speech, but only direct and public incitement to commit genocide is prohibited under Article 2(3)(c) of the Statute. This conclusion is corroborated by the travaux preparatoires to the Genocide Convention.").

48. *Id*. § 692.

49. *See generally* Prosecutor v. Šešelj, Case No. ICTY-03-67-T, Trial Judgement (Mar. 31, 2016), https://www.icty.org/x/cases/seselj/tjug/en/160331.pdf.

50. *See generally* Prosecutor v. Šešelj, Case No. MICT-16-99-A, Appeal Judgement (Apr. 11, 2018), https://ucr.irmct.org/LegalRef/CMSDocStore/Public/English/Judgement/NotIndexable/MICT-16-99-A/JUD282R0000519025.pdf.

51. *See, e.g., ICCPR supra note 26; see also https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/SeminarRabat/Rabat_threshold_test.pdf.*https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/SeminarRabat/Rabat_threshold_test.pdf.

52. ICCPR, supra note 26 at § 19(2)-20(2).

53. Article 38 of the Statute of the International Court of Justice identifies the following sources of international law: (i) international conventions; (ii) international custom, as evidence of a general practice accepted as law; (iii) general principles of law recognized by nations; and (iv) judicial decisions and teachings of the most highly qualified publicists, as subsidiary means for the determination of the rules of law. Statute of the International Court of Justice, Art. 38, para. 1.

> The Commission shall draw inspiration from international law on human and peoples' rights, particularly from the provisions of various African instruments on human and peoples' rights, the Charter of the United Nations, the Charter of the Organization of African Unity, the Universal Declaration of Human Rights, other instruments adopted by the United Nations and by African countries in the field of human and peoples' rights as well as from the provisions of various instruments adopted within the Specialized Agencies of the United Nations of which the parties to the present Charter are members.[54]

Article 61 of The African Charter similarly states:

> The Commission shall also take into consideration, as subsidiary measures to determine the principles of law, other general or special international conventions, laying down rules expressly recognized by member states of the Organization of African Unity, African practices consistent with international norms on human and people's rights, customs generally accepted as law, general principles of law recognized by African states as well as legal precedents and doctrine.[55]

The ACHPR and the African Court require that local remedies must be exhausted before a matter is brought before them; local remedies refer to any judicial or legal mechanisms put in place at the domestic level to ensure the effective settlement of disputes. Generally, the matter must be brought before the highest appellate court for a decision.[56] Therefore, in practice, international law's preference for local remedies means that most human rights cases should be brought at the domestic level first.

Looking to domestic hate speech specifically, many nations across the world have established legal limitations on hate speech. Many of these laws came about in the wake of World War II, and were designed to curb incitement to racial, ethnic, and religious hatred after the Holocaust. For example, in Germany, it is illegal to publicly incite hatred against parts of the population, to call for violent or arbitrary measures against them, or to insult, maliciously slur, or defame them in a manner violating their human dignity.[57] In 2017, Germans criminalized hate speech on social media sites, imposing large fines for platforms failing to remove illegal content.[58] The French Penal code and press laws similarly prohibit communication that is defamatory or

54. *African Charter, supra note 35, at § 61.*

55. *Id.* at Art. 61.

56. SINGH, *supra* note 29, at 34.

57. Strafgesetzbuch [STGB] [Criminal Code], § 130 https://www.gesetze-im-internet.de/stgb/__130.html.

58. Netzwerkdurchsetzungsgesetz [NetzDG] [Network Enforcement Act] Oct. 1, 2017, BGBL. I (Ger.).

insulting, or that incites discrimination, hatred, or violence against a person or group based on specific criteria.[59]

Drawing from its history of apartheid, South Africa has one of the world's most detailed and comprehensive laws against hate speech, accounting for groups and attributes absent from many other countries' laws, such as pregnancy, marital status, conscience, language, color, and "any other group where discrimination… causes or perpetuates systemic disadvantage; (ii) undermines human dignity; or (iii) adversely affects the equal enjoyment of a person's rights and freedoms."[60]

The global outlier in treatment of hate speech is the United States. In America, hate speech is protected under the rubric of the First Amendment, though it is not expressly defined.[61] This is important because the dominant social media companies—like Facebook, Twitter, Instagram, and WhatsApp—are American-based. While this dominance may be shifting with the rise of TikTok,[62]—a Chinese-owned social media company part of the ByteDance conglomerate—popular sentiment in Silicon Valley roots online speech in the *marketplace of ideas* and promotes "more speech" as a remedy for abhorrent speech.[63] Still, this does not mean that anything goes under American jurisprudence. Under the First Amendment, hate speech is not protected when it meets certain exceptions for protected speech, such as directly inciting imminent criminal activity[64] or consisting of specific threats of violence targeted at a person or group.[65]

Social media platforms are not bound by the First Amendment and create their own terms of service for self-governance.[66] However, they may still be bound to national laws that mandate or sanction speech-related activity in other countries. Scholars have referred to these tech companies as "the new governors" for the power and broad scope of influence these platforms possess, which rival the States of populations served in the geopolitical impact of their internal rules.[67]

---

59. Nicolas Boring, *Limits on Freedom of Expression*, Library of Congress (2019), https://www.loc.gov/law/help/freedom-expression/france.php.

60. Qwelane v. South African Human Rights Commission, 7 (2019), http://globalfreedomofexpression.columbia.edu/cases/qwelane-v-south-african-human-rights-commission.

61. U.S. CONST. amend. I.

62. *About TikTok*, TIKTOK, https://www.tiktok.com/about?lang=en.

63. *Two Models of the Right to Not Speak*, 133 HARV. L. REV. 2359, 2372-73 (2020).

64. Brandenburg v. Ohio, 395 U.S. 444, 447 (1969).

65. Watts v. United States, 394 U.S. 705, 707 (1969).

66. Public opinion around modifying Section 230 of the Communications Decency Act of 1996 is gaining momentum. Some have proposed removing or limiting intermediary liability protections, to modify platform's engagement with the content they host. *See generally* JEFF KOSSEFF, TWENTY-SIX WORDS THAT CREATED THE INTERNET (2019).

67. *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018).

C. *Dangerous Speech and Hate Speech*

Hate speech is generally understood as a form of speech intended to vilify or disparage a person or group of people based on their membership in an identity group, with the potential to bring them harm.[68] Understandings of what constitutes intention, vilification or disparagement, relevant identity groups (often called "protected categories"), and relevant kinds of harm vary widely and are highly dependent on audience and local context.[69]

One proposal for a narrower subset of hate speech comes from Susan Benesch, in her canonical theory of "dangerous speech." As previously alluded to, Benesch defines this as "[a]ny form of expression (e.g. speech, text, or images) that can increase the risk that its audience will condone or commit violence against members of another group."[70] Although the concepts necessarily overlap, dangerous speech is distinct from hate speech in that it specifically focuses on language's "…capacity to inspire a harm that is all too easy to identify—mass violence—and that almost everyone can agree on wanting to prevent;"[71] "hate speech" relies on an emotional component—hatred—which is subjective and can be difficult to measure.

The theory of dangerous speech has been very influential in articulating a clear definition of speech that rises to the threshold of incitement of physical violence, and especially in identifying hallmarks of this type of speech, which recur through history and across cultures. My research with communities in Kasese, Uganda leans heavily on the dangerous speech framework, both in the material presented to community partners at workshops and the parameters applied in data analysis with U.C. Berkeley's Human Rights Center.[72] As the research results will demonstrate, Uganda provides a prime example of postcolonial hate speech, where online spaces cannot be viewed through an ahistorical lens.

---

68. Benesch & Maynard, *supra* note 5.

69. *See, e.g.*, *UN Actions Against Hate Speech*, UNITED NATIONS, https://www.un.org/en/hate-speech/united-nations-and-hate-speech/international-human-rights-law.

70. *Dangerous Speech: A Practical Guide*, DANGEROUS SPEECH PROJECT (2021), https://dangerousspeech.org/guide.

71. *Id*.

72. This article primarily focuses on hate speech (as opposed to limiting the project's scope to dangerous speech) because of the ubiquity of the term "hate speech" in the international conversation around content moderation. Additionally, "hate speech" resonated more immediately with the local partners in Africa. Furthermore, "hate speech" is the term used by Facebook, the dominant platform in Kasese, Uganda. The research objective was to develop an understanding of hate speech that was grounded in a local context, rather than to create a rigid definition to be used in content moderation, with hope that the former may be able to inform the latter.

D. *Ugandan Politics and Law*

1. Colonial Past

In addition to understanding theories of hate speech, we must understand local history to gain a fuller contextual understanding of Ugandan hate speech. Uganda is a country in eastern Africa approximately the size of Great Britain.[73] The Kasese District, located in the far southwest, shares a border with the Democratic Republic of the Congo.[74] This region's past is vital to explaining the current social, ethnic, religious, and tribal strife that still plagues Kasese today.

Long-standing political tensions in Uganda can be traced to pre-colonial land disputes.[75] These disputes were compounded after the British Protectorate of Uganda was established in 1894, establishing new British borders that did not consider the preexisting governance, history, or autonomy of the people already living there.[76]

The British colonial rulers deliberately created stratifications within the Ugandan territory.[77] The Under this "divide-and-rule" strategy, economic power and education was concentrated in the south, but the north provided the military power.[78] As a result, southerners occupied academic, judicial, bureaucratic, and religious positions and became seen as social elites.[79] Police and armed forces came from northern tribes. The name "Uganda" comes from one favored tribe's name, the Buganda, who still enjoyed semi-autonomy under British rule as part of their favored status.[80]

The impacts of these divisions shaped the political events of postcolonial Uganda. The country gained its independence on October 9, 1962.[81] This began a series of power struggles between the kingdoms and the centralized government.[82] Governance was marked by confusion. Uganda was a quasi-federal State with five regional monarchies, non-monarchical districts, and a central government until 1967.[83] At that time, the government adopted a

73. Omari. H. Kokole, Kenneth Ingham, Maryinez Lyones & M. Semakula M. Kiwanuka, *Uganda*, ENCYCLOPEDIA BRITANNICA (Oct. 19, 2022), https://www.britannica.com/place/Uganda.

74. *Kasese, Uganda*, GOOGLE MAPS, https://www.google.com/maps/place/Kasese,+Uganda/@0.1183366,29.722389510z/data=!3m1!4b1!4m5!3m4!1s0x1761f0681ecdc3b9:0xaf31ee3aa62d09c3!8m2!3d0.0646285!4d30.0665236.

75. Kokole, *supra* note 73.

76. Interview with Johncation Muhindo, Kasese, Uganda (Sept. 12, 2019).

77. *See* Kokole, *supra* note 73.

78. *Id.*

79. *Id*.

80. GODFREY MWAKIKAGILE, UGANDA: A NATION IN TRANSITION: POST-COLONIAL ANALYSIS, 28–29, 34 (2012).

81. *See* Kokole, *supra* note 73.

82. *See id.* para. 102-06.

83. *Id.* para. 56.

constitution that abolished the monarchies and gave political power to an elected president.[84] As a result, regions that had a strong kingdom and had desired full autonomy, like the Kasese region, found themselves fighting against the central government for recognition and authority.[85]

Following Ugandan independence, a series of violent political transitions further exacerbated tribal and geographic tensions.[86] Continued disputes over the authority of tribe-specific kingdoms in relationship to the Ugandan government have fueled internal intertribal conflict.[87] This pattern has resulted in extreme political polarization.[88] In this atmosphere, political identity (rather than ethnicity) and tribal identity are intricately interwoven.[89] The administrations of Idi Amin and Milton Obote further amplified political tensions and regional reactions to a strong centralized state they did not feel represented their interests, or even favored their rivals.[90] Secessionist movements surged, especially in the area around Kasese.[91]

2. Recent Violence

Over the past few years, Kasese has experienced outbreaks of ethnic violence closely related to national and electoral politics. In November 2014, Human Rights Watch released a report detailing instances of deadly ethnic violence and reprisals over the autonomy of tribal cultural institutions[92] which left over ninety-two dead.[93] Between February and April of 2016, nearly fifty people were killed during political skirmishes regarding disputed election seats.[94] In November 2016, a massacre perpetrated by national security forces

84. *Id.*

85. *See id.* para. 102-06, 114-15.

86. *Id.* para. 106-07, 109-12.

87. *See id.* para. 106, 113.

88. *See id.* para. 102-07, 109-13.

89. *See infra* Section III.D (showing that this was one of the most surprising outcomes of the Kasese study, as most online platforms presume ethnicity—and not political identity—is aligned with tribal identity).

90. *See* Kokole, *supra* note 73, para. 106-07, 111-13.

91. *Why NRM Lost Kasese District Vote*, MONITOR (MAR. 6, 2016), https://www.monitor.co.ug/Elections/NRM-lost-Kasese-District-vote/2787154-3125692-ngew85/index.html.

92. In 2014, President Museveni recognized the autonomy of the Bamba Kingdom separate from both the Tooro and Rwenzururu Kingdoms. Although there are three separate cultural institutions for these tribes, because of the histories of migration and conflict, individuals from all these three tribes live coexist with one another in the geographic territory of each kingdom.

93. *Uganda: Violence, Reprisals in Western Region*, HUMAN RIGHTS WATCH (Nov. 11, 2014, 1:00 AM), https://www.hrw.org/news/2014/11/05/uganda-violence-reprisals-western-region.

94. *See* Anna Reuss & Kristof Titeca, *There is new violence in Western Uganda. Here's why.*, WASHINGTON POST (Nov. 29, 2016, 11:00 AM), https://www.washingtonpost.com/news/monkey-cage/wp/2016/11/29/what-is-happening-in-uganda/; *see also Uganda: No Justice for 2016 Kasese Massacre by Security Forces*, HUMAN RIGHTS WATCH (Oct, 10, 2018, 4:00 AM),

left over one hundred dead at the Rwenzururu palace compound in Kasese, and political opposition leaders were imprisoned.[95]

In light of the relationship between online hate speech and offline violence, activists in Kasese raised concerns that online discourse may have helped instigate the recent ethnic violence and may contribute to future clashes, especially surrounding elections.[96]

This dynamic of politicization and tribal tension plays out in online spaces. Social media has gained massive traction in Uganda, especially among wealthier citizens and social elites: out of Uganda's population of approximately 42.7 million, 10.67 million have internet access, and 2.5 million are active social media users.[97] In recent years, social media has been widely used as a medium for political mobilization in Uganda, enabling the ascent of anti-establishment politicians such as People Power candidate and popular musician Bobi Wine.[98] However, as this research will show, social media in Uganda has also become an avenue for divisive messages that may incite violence along social fissures.

While the Ugandan government has not publicly recognized the threat of hate speech, President Yoweri Museveni has addressed the phenomenon in denouncing disinformation and "gossip."[99] In February 2016, immediately before the presidential election, and again in May 2016, in the days leading up to his inauguration, Museveni imposed internet shutdowns[100] under the guise of controlling the spread of "fake news" and "gossip."[101] This was a ruse, according to critics who argued that controlling information was intended to suppress unpopular sentiment over his government.[102] In May 2018,

---

https://www.hrw.org/news/2018/10/10/uganda-no-justice-2016-kasese-massacre-security-forces.

95. *See Uganda: No Justice for 2016 Kasese Massacre by Security Force*, *supra* note 94.

96. While the initial plan was to monitor the impact of this study on the 2021 presidential elections, it was not possible to return to Kasese because of the global pandemic. Furthermore, the Museveni government later banned Facebook from Uganda. More research should be done about the long-term impact of hate speech interventions, especially in areas experiencing a history of continuous ethnic conflict.

97. Simon Kemp, *Digital 2020: Uganda*, DATAREPORTAL (Feb. 18, 2020), https://datareportal.com/reports/digital-2020-uganda.

98. *See* Bobi Wine, WIKIPEDIA, https://en.wikipedia.org/wiki/Bobi_Wine (last updated Dec. 15, 2022).

99. *Uganda imposes WhatsApp and Facebook tax 'to stop gossip,'* BBC NEWS (May 31, 2018), https://www.bbc.com/news/world-africa-44315675.

100. Berhan Taye, *Time is up: Uganda in court over internet shutdowns that violate human rights*, ACCESS NOW (Nov. 8, 2018, 4:42 PM).

101. *Id*.

102. Hilary Smith & Jeffrey Matfess, *Africa's Attack on Internet Freedom*, FOREIGN POLICY (July 13, 2018), https://foreignpolicy.com/2018/07/13/africas-attack-on-internet-freedom-uganda-tanzania-ethiopia-museveni-protests/.

Museveni also imposed a social media tax,[103] widely criticized as an effort to curb free speech.

Although Ugandan law makes no mention of hate speech, it contains provisions designed to curb critical online content and penalize incitement to violence. Most of these provisions come from the Museveni regime, invoking counterterrorism and safety as a cover to control public debate and dampen rising discontent and an increasingly popular opposition.[104] According to a press release issued by the Uganda Police Force in August 2018, "[h]ate speech and messaging contravenes [1] Computer Misuse Act 2012 Section 25 and [2] The Penal Code Act Section 51(1) (a & b) incitement to violence."[105] The Computer Misuse Act 2011, Section 25, states:

> Any person who willfully and repeatedly uses electronic communication to disturb or attempts to disturb the peace, quiet or right of privacy of any person with no purpose of legitimate communication whether or not a conversation ensues commits a misdemeanor and is liable on conviction to a fine not exceeding twenty four currency points or imprisonment not exceeding one year or both.[106]

The Penal Code Act, Section 51(1)(a)-(b), similarly states:

> 1. Any person who without lawful excuse, prints, publishes or to any assembly makes any statements indicating or implying that it would be incumbent or desirable—(a) to do any acts calculated to bring death or physical injury to any person or to any class or community of persons; or (b) to do any acts calculated to lead to destruction or damage to any property, commits an offence and is liable to imprisonment for three years.[107]

At least two other provisions of the Penal Code Act address potential acts of hate speech. Section 83(1) on "Incitement to violence" states:

> (1) Any person who incites any other person to do an act of violence against any person by reason of his or her race, place of origin, political opinions, colour, creed or sex or office commits an offence and is liable on conviction to imprisonment for a term not exceeding fourteen years.[108]

---

103. Juliet Nanfuka, *Uganda Blocks Access to Social Media, VPNs and Dating Sites as New Tax Takes Effect*, CIPESA News Blog (July 1, 2018), https://cipesa.org/2018/07/uganda-blocks-access-to-social-media-vpns-and-dating-sites-as-new-tax-takes-effect/.

104. *See supra* notes 99-103.

105. Wilfred Kamusiime, *Police warns on spread hate speech messages*, UPF PRESS RELEASE (Aug. 8, 2018), https://www.upf.go.ug/police-warns-on-spread-hate-speech-messages/.

106. Computer Misuse Act § 2, 2011 (Uganda).

107. Penal Code Act § 120, 1950 (Uganda).

108. *Id.* at § 83 (1).

Section 41 of the Penal Code Act addresses "Promoting sectarianism":

> (1) A person who prints, publishes, makes or utters any statement or does any act which is likely to— (a) degrade, revile or expose to hatred or contempt; (b) create alienation or despondency of; (c) raise discontent or disaffection among; or (d) promote, in any other way, feelings of ill will or hostility among or against, any group or body of persons on account of religion, tribe or ethnic or regional origin commits an offence and is liable on conviction to imprisonment for a term not exceeding five years.
>
> (2) It shall be a defence to a charge under subsection (1) if the statement was printed, published, made or uttered, or the act was done with a view to exposing, discouraging or eliminating matters which promote or have a tendency to promote sectarianism.

The Ugandan law alluded to by the president bans the promotion of sectarian hate, and therefore this might have been the law that Museveni invoked in calling for arrests.[109] Museveni did not specify the exact law he was using, but political maneuvering like this is one way governments abuse laws that criminalize libel to crack down on political dissidents.[110] Analysis from Freedom House supports this concern, warning about the negative, overbroad stifling of free speech under the guise of policing online hate speech:

> The penal code contains provisions on criminal libel and the promotion of sectarianism, imposing penalties that entail lengthy jail terms. While none of these laws contain specific provisions on online modes of expression, they could arguably be invoked for digital communications and generally create a 'chilling effect' on freedom of expression both online and offline…[I]n May 2015, President Yoweri Museveni called for the arrest of three individuals whose voices were heard in an audio clip disseminated widely via WhatsApp that reportedly contained abusive and sectarian language. According to observers, the government's increased crackdown on online speech, particularly on social media platforms, in the past year may be an indication of restrictions to come in the lead up to the presidential election in 2016."[111]

Regulatory bodies in Uganda have also been part of enforcement against hate speech, with the potential attendant peril of criminal regulation on

109. Lizabeth Paulat, *Uganda President Challenges Free Speech on Social Media*, VOICE OF AMERICA (June 4, 2015), https://www.voanews.com/a/uganda-president-challenges-free-speech-on-social-media/2807512.html

110. *See, e.g., Uganda,* FREEDOM ON THE NET 2015 (Oct. 28, 2015), https://freedomhouse.org/sites/default/files/resources/FOTN%202015_Uganda.pdf.

111. *Id*.

freedom of speech. The Uganda Communications Commission (UCC) "noted with concern the increasing incidences where broadcasters offer platrfom [sic] to members of the public to express views that are inciting, discriminating and stirring up hatred or violence," and "remind[ed] broadcasters that they have an obligation to sieve content likely to cause undue offence and breach minimum broadcasting standards."[112] The purpose of the notice was "to warn and remind all broadcasters to strictly comply with their statutory and license obligations, failure of which will leave UCC with no alternative but to involve regulatory sanctions under Section 41 of the Uganda Communications Act and/or institute criminal proceedings against offending broadcasters at their own peril."[113]

Again, overinclusive laws may be used to stifle freedom of speech, through criminal or administrative sanctions, if those in power determine the message of a broadcast is akin to hate speech.

II. METHODOLOGY AND STUDY DESIGN

A. *Workshops and a Community-Generated Definition of Hate Speech*

In light of the fear by civil society over repeat violence in upcoming elections, the research team created an approach to tease out how hate speech was operating in Kasese, Uganda. We examined linguistic and computational data and worked with a local team of community experts to analyze our results. In practice, this involved creating community-generated definition of hate speech and generating a crowdsourced lexicon of hate speech terms.

This project was initiated by Ugandan human rights activists, who proposed partnering with topical experts on hate speech in Kasese, Uganda to institute a series of on-the-ground workshops in September 2019.[114] My U.C. Berkeley-based research team began by developing a methodology to define and capture hateful speech. With the help of Peace Tech Lab, we created a

112. Paul Ampurire, *UCC Warns Broadcasters Against Facilitating Hate Speech*, SOFTPOWER NEWS (June 13, 2018), https://www.softpower.ug/ucc-warns-broadcasters-against-facilitating-hate-speech.

113. *Id*.

114. The project was made possible through the time, expertise, and generosity of several partners. It was funded through a Mandela Washington Fellowship award from the U.S. Department of State. Musoki Elizabeth, an Africa Mandela Washington Fellow, proposed the initial partnership to me and she was vital to the success of the work in Kasese. Local Ugandan non-profit professionals, like Johncation Muhindo, made project work possible. Foley Hoag LLP provided pro bono support for the legal aspects of the work. The Peace Tech Lab provided insight and a foundation for the study methodology. Sonnet Phelps provided on-the-ground research, with the guidance and support of Alexa Koenig and Andrea Lampros, the project team at Berkeley Human Rights Center (HRC). Subsequent open-source data analysis was also conducted in partnership with the HRC.

curriculum to educate our local audience about the theory and impact of hate speech.[115] Local workshop participants were identified by a team of local partners in Kasese for their diversity and their influence in the community.

Participants were trained on international perspectives on hate speech and were then invited to voice perspectives that might better capture and reflect local understandings. Local civic, religious, political, and community leaders were grouped into mixed cohorts to develop their own definition of hate speech, applying their understanding of local events, social dynamics, and nuances of speech to the curriculum. Each proposal was reviewed with all workshop participants and discussed.

After that portion of the multi-day workshop, we established a consensus-based definition of hate speech derived from the community output:

> Hate speech is [speech that is] not protected by freedom of expression. It is a communication, targeting either a community or a representative of a community. Hate speech frequently degrades, incites violence against, or promotes negative stereotypes about its targets.[116]
>
> Hate speech is always reliant on collective knowledge. The history and shared cultural understanding of a community determine the context that makes a negative message into hate speech.
>
> A community can be characterized by shared traits, including religion, sexual orientation, tribal or political affiliation, ethnicity, social class, economic group, educational level, age, gender, disability, tradition, language, national origin, or geography.

This community-generated definition, with its capacious definition of harm, its emphasis on collective knowledge, and its long list of protected classes, is much broader and more open to interpretation than even international standards, as will be discussed later in this paper. This is by the drafters' design: the community wanted the definition to be broad enough to integrate most participants' perspectives. Similarly, a related distinction is the use of the word "community," where other definitions use "identity group" or "protected category" because it resonated most strongly with the participants and indicated that some at-risk characteristics for targeting were voluntary or chosen.

---

115. The methodology to create a lexicon of local hate speech was adapted through the kindness and expertise of Peace Tech Labs. Their staff has developed a resource to generate hate speech lexicons from local communities. This project graciously borrowed some of their insights and methods, but applied them in a more constrained timeline, and provided feedback back to Peace Tech Labs.

116. *See* Chapter 120 Uganda Penal Code Act § 83(1) (stating that "incitement to violence" is a criminal act).

Additionally, this definition lists three types of potential harm to targets of hate speech—degradation, incitement to violence, and the promotion of negative stereotypes—but the inclusion of "frequently" leaves the possibility that it may produce other forms of harm not listed.

Furthermore, the assertion that "hate speech is always reliant on collective knowledge" underscores the importance of local context. It acknowledges that a "negative message" becomes hate speech on the basis of "history and shared cultural understanding of a community." Participants also noted that "history" is intended to refer to Uganda's colonial history and the legacy of colonialism in the present, as well as more recent conflicts and political transitions.

The inclusion of "tradition" is in line with some of the distinctive characteristics of Africa-centric treaties and declarations in international law. As will be further elaborated, the participants' inclusion of history, cultural understanding, and tradition emphasizes the importance of including a parallel track of context when analyzing hateful speech, when the dominant national narrative and legacy political structure is closely aligned with colonizers.

The participants and I weighed each of the listed "shared traits" carefully and determined that despite some overlap (e.g., between "social class" and "economic group" or "tradition" and "tribal affiliation"), the community preferred to include all of the categories that participants identified.

B. *Lexicon and Open-Source Data Collection*

The next step was applying this definition to generate a lexicon of slurs, crowdsourced from community experience, at the end of the workshops. The investigators and I took care to create new and diverse local groups. We created both verbal and written means for responses, to capture terminology and opinions of participants who may not have wished to speak out loud or in front of a group.

After organizing and reviewing the initial list of hate speech terms, we held several sessions with our Ugandan human rights experts to clarify the context where investigators needed more information on the meaning, context, history and use of the epithets. Local experts approved the final hate speech lexicon.

Upon completing fieldwork, the U.C Berkeley Human Rights Center team conducted an open-source investigation of Ugandan social media that resulted in a database of 102 pieces of content collected from twenty-seven public Facebook groups.[117] Using the lexicon, the research team then collected content that (upon first review) met the threshold for hate speech based on the community-vetted specific criteria. We then engaged in a second classifying process to confirm whether the content met our hate speech

117. *See infra* App. I.

parameters, including explaining what category of hate speech an epithet belonged to and what group it targeted.

The content we collected was highly context-dependent and required a significant understanding of local political, tribal, and cultural dynamics for proper analysis and classification. We realized that even with tutorials, our education on the politics, geography, and history of Kasese was insufficient to understand some of the linguistically and contextually nuanced content, which would only be discernable by someone with deep local knowledge. For example, the multiple languages present in Kasese made translating and contextualizing non-English words and phrases difficult. Similarly, context was challenging to interpret, especially when terms often associated with hate speech were used in a benign context. Additionally, distinguishing satirical political humor from legitimate calls for violence was not a straightforward exercise. While contextual challenges present difficulties in parsing any form of hate speech, the phenomenon was especially pronounced in Kasese's post-colonial context.

Overall, local stakeholder involvement in our collection and evaluation processes was essential for hate speech analysis. We mitigated our cultural and linguistic limitations by engaging in a partner-led review process with our Kasese community activists and human rights experts, who provided linguistic and contextual corrections, and evaluations of our data and preliminary results. This feedback was integrated into our dataset and findings. As a result, we consider our experience to be exploratory, rather than definitive.

During the classification process, our team's difficulties were similar to those experienced by social media platforms in creating and enforcing community standards. The material proved challenging to analyze and categorize. Judging whether a particular piece of content constituted hate speech rarely, if ever, was simple.

III. RESULTS

A. *Online Hate Speech*

Applying the Kasese community definition of hate speech, fifty-three pieces of content were classified as hate speech. Of the content identified as hate speech, thirty-seven targeted people on the basis of tribal affiliation, with the most frequent targets being members of the Bakonzo (sixteen), Batooro (six), and Baganda (five) tribes.[118] Nine of the posts targeted women on the basis of their gender. The three posts targeting individuals on the basis of religion were directed at Muslims.

118. These statistics are not intended to be fully representative of the discourse, as our team's inquiry was exploratory rather than exhaustive.

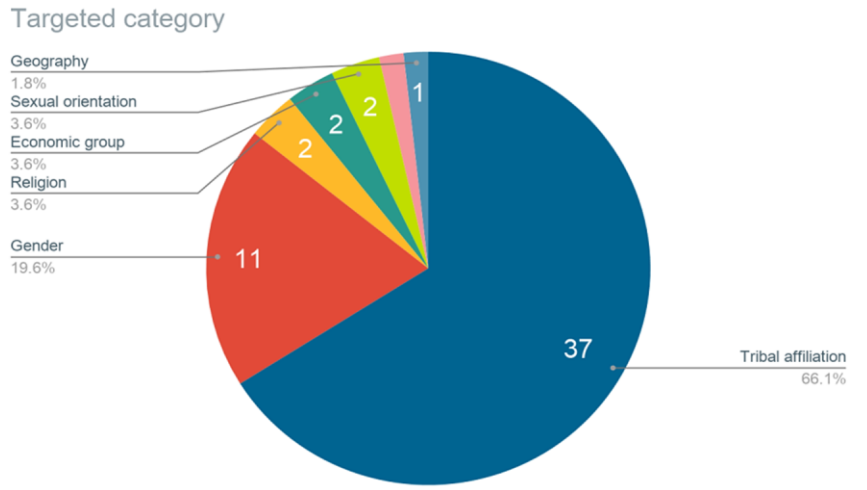

FIGURE 1. DISTRIBUTION OF HATE SPEECH ACROSS TARGETED CATEGORIES[119]

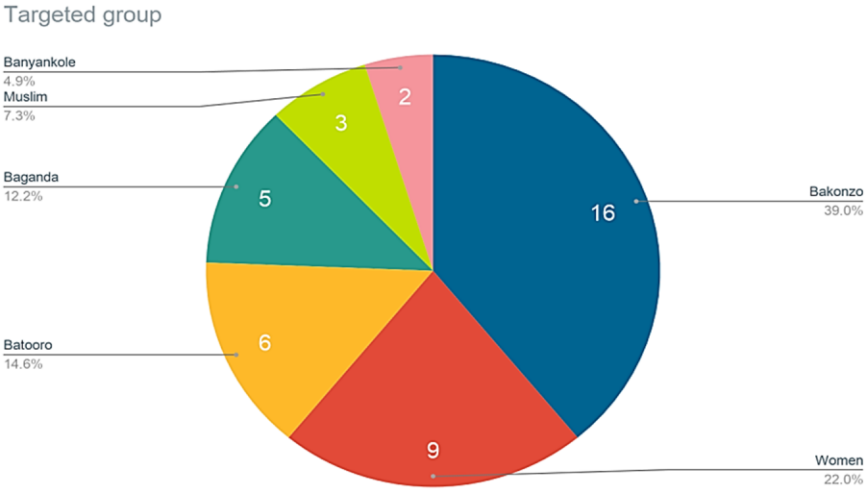

FIGURE 2. DISTRIBUTION OF HATEFUL SPEECH BY TARGETED GROUP

The five hallmarks of dangerous speech were used to better analyze the character of the information—with an important new addition. As will be described later in this paper, the five canonical hallmarks are characteristics of hateful speech that incites violence, which can be traced across societies, time

119. All figures were created by the U.C. Berkeley Human Rights Center and are reprinted here with their permission.

periods, and cultures:[120] these are dehumanization, reinforcing harmful cultural stereotypes, threats to group integrity or purity, assertions of attacks against women and girls, and projection of perpetrators' threats and their harmful intentions onto intended victims, also known as "accusation in a mirror."[121]

The data reflected a trend that had been present in the workshops. Participants described as "hateful" content that disputed specific tribes' right to be on specific areas of land or called for certain groups' removal or banishment. Terms in our lexicon reinforced this finding: *ekithaka nikyethu,* meaning "the land is ours"; *abalihanda*, meaning "outsiders" or "other tribes"; and *omulihanda*, meaning "enemy" usually "of the enemy tribe."[122] Content that geographically excluded surfaced frequently during our research: out of the content that was coded as hate speech, 22% of posts called for geographic exclusion. This category surfaced as the third most frequent category of speech in the dataset.[123]

Including a geographic exclusion metric allowed for an examination of hate speech that called for the banishment or expulsion of certain groups from areas of land as distinct from national origin or ethnicity. However, the relevance and prevalence of this type of intertribal hate speech was a distinctive factor on its own.

While this is an initial study, it demonstrates what may happen if online spaces are considered outside their historical contexts. The phenomenon taps into the still-present tensions that originated in the postcolonial legacy of intertribal conflicts over territory and political control. For example, *endaghangali* means "traitor" and was explained with the following context: "It was during the 1962–1982 Rwenzururu struggle. If you were called such, it means you betrayed the cause and sided with the enemy against your own."[124]

In other words, the hate speech found in our study looked back to British colonial legacies involving putting tribes competing for land in the same political units, while establishing tribal "kings" as mere cultural leaders without any political or administrative power to govern.[125] There was confusion in having the king serve as a figurehead instead of the head of a functional monarchy.[126] Events in Ugandan history showed up like ripples from a stone dropped into a well.

---

120. Dangerous Speech Project, *supra* note 70.

121. *Id*.

122. *See infra* App. I.

123. In this analysis, seventeen posts reinforced harmful cultural stereotypes, fourteen involved dehumanization, twelve included calls for geographic exclusion, seven described threats to group integrity or purity, four were accusations of attacks against women or girls, and one was an accusation in a mirror.

124. Interview with Musoki Elizabeth, Kasese, Uganda (Sept. 12, 2019).

125. GODFREY MWAKIKAGILE, UGANDA: A NATION IN TRANSITION: POST-COLONIAL ANALYSIS 18 (New Africa Press, 2012).

126. *Id*. at 7.

As such, I propose to add a sixth category, "calls for geographic exclusion," as a new hallmark of dangerous speech. Kasese's example demonstrates how postcolonial legacies prime this type of content to provoke violence: Ugandan divisions between tribal kingdoms and other traditional centers of power; social animosity fueled by the British between northern and southern groups; the division of military centers of power in the north and other elites in the south, resulting in turbulence and dictatorships; the British forcefully combining formerly independent kingdoms to create Uganda, and then denying the Baganda self-rule after colonial withdrawal; the division between separatist regions like the Kasese district and the affiliation of those viewpoints with certain tribes.[127]

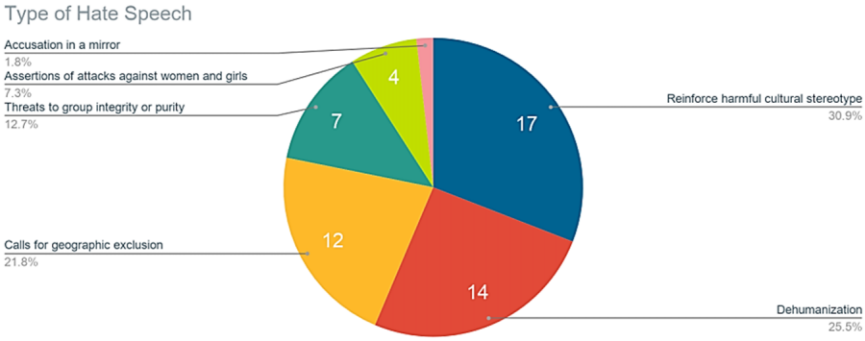

FIGURE 3.DISTRIBUTION OF HALLMARKS OF DANGEROUS SPEECH

B. *Local Hate Speech*

Using online data analysis and qualitative interviews reflects the local community's lived experience but does not represent an exhaustive sample of hate speech in the region. However, the open-source social media analysis provided evidence to support community leaders' concern that online discourse may play a role in intensifying local conflict. The examples below demonstrate the necessity of local context in parsing hate speech. In doing so, a portrait of the local tensions in Kasese emerges.

127. *Id.*

1. Religion

FIGURE 4. RELIGION SOCIAL MEDIA EXAMPLE

Religion is a dominant force in Kasese. Local leaders identified diverse religions present in Kasese, including Anglican, Christian, Protestant, Muslim, Catholic, Seventh-Day Adventist, Pentecostal, and Baptist faiths.[128]

After independence from Britain, the first Ugandan government emerged alongside regional blocs and interest groups—including the Protestant (predominantly Anglican), Catholic, and Islamic contingents. These factions lacked a common agenda or plan for national unity.[129] As such, religious groups became a synecdoche for different tribes and political interests. Hate terms we encountered included epithets stemming from religious culture, like *omutsule* and abat[s]ule, meaning "uncircumcised."[130]

Most content relating to religion that was flagged as hate speech or potential hate speech targeted Muslims. This group comprises about 1/8 the population of Uganda and is the third most popular religion, behind Protestantism and Catholicism.[131] Such content uses a stereotype conflating Muslims and terrorism to "other" that substantial minority.

128. Interview with Johncation Muhindo, *supra* note 76.
129. MWAKIKAGILE, *supra* note 125, at 11-12.
130. *See infra* App. I.
131. *Religion of Uganda*, ENCYCLOPEDIA BRITANNICA, https://www.britannica.com/place/Uganda/Religion (last visited Aug. 18, 2022).

2. Sexual Orientation

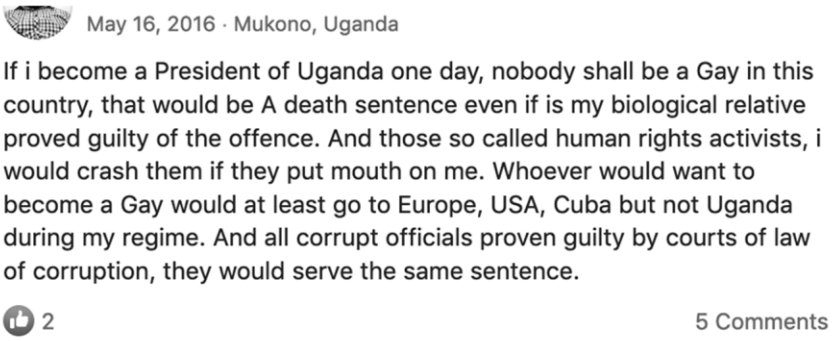

FIGURE 5. SEXUAL ORIENTATION SOCIAL MEDIA EXAMPLE

Homosexual activity is illegal in Uganda.[132] Past legislation has proposed enforcing related laws with the death penalty.[133] Sexual orientation is a common topic in public discourse, but often includes homophobic language and intent. Examples of hate speech stemming from the community lexicon included *ebisiyaga,* meaning "homosexual" as a derogatory term. Given the illegal status of homosexual acts, accusations of homosexuality may qualify as dangerous speech by representing a sublimated call for violence or vigilante enforcement of the law.[134]

132. The Anti-Homosexuality Bill, 2009 (Bill No. 18/ 2009) (Uganda).

133. Nita Bhalla, *Uganda Plans Bill Imposing Death penalty for Gay Sex*, REUTERS, Oct. 10, 2019, https://www.reuters.com/article/us-uganda-lgbt-rights/uganda-plans-bill-imposing-death-penalty-for-gay-sex-idUSKBN1WP1GN.

134. An analogous example would be calling immigrants "illegals" to allude to their lacking legal status.

3. Tribal Affiliation

FIGURE 6. *TRIBAL AFFILIATION* SOCIAL MEDIA EXAMPLE

Tribes in Uganda are less closely aligned with ethnicities—and according to a community member, perhaps better understood like family houses in Game of Thrones. They are more political entities than ethnic groups, according the Kasese community and scholars.[135] A person can marry into a tribe, which changes their affiliation, regardless of what tribe they were born into.

Within the South, tribes share a Bantu linguistic heritage. Bantu speakers form the largest portion of Uganda's population. Of these, Baganda remain the largest single group, constituting roughly one-sixth of the total national population. Other Bantu speakers are the Basoga, Bagwere, Bagisu, Banyole, Basamia, Batooro, Banyoro, Bakiga, Banyankole, Bamba, and Bakonjo.[136]

Common tribal affiliations in Kasese are Bakonjo, Batooro, and Basongora.[137]

Tribal affiliation-based tensions have root in colonial governance structures. British administrators placed people of different tribal affiliations near each other, within the same Kingdoms in Uganda.[138] This deliberate strategy pitted tribes against each other for access to resources, administrative

135. Interview with Musoki Elizabeth, *supra* note 124.

136. *Uganda Culture*, TRUST INVESTMENT GROUP, https://trustinvestmentsgroup.com/national-parks/# (last visited Feb. 4, 2023); *see also Bantu Peoples*, ENCYCLOPEDIA BRITANNICA, https://www.britannica.com/topic/Bantu-peoples (last visited Aug. 18, 2022).

137. *Background*, KASESE DISTRICT LOCAL GOVERNMENT, https://www.kasese.go.ug/about-us/background/. Interview with Musoki Elizabeth, *supra* note 124.

138. *See* Kokole, *supra* note 73.

positions, and opportunities for education, wealth, and advancement.[139] Scarcity and competition, along with favoritism, created and exacerbated tensions among tribal groups. It is not surprising that a high percentage of discovered content perpetuated tribal stereotypes and exemplified intertribal tensions.

Terms in the lexicon reinforced this analysis. *Abanyagwagwa* means "foreigners" but was not applied to people outside Uganda; instead, it was primarily used to refer to other tribal groups.

4. Ethnicity

FIGURE 7. ETHNICITY SOCIAL MEDIA EXAMPLE

Ethnic groups differ from tribes, in that they are fixed throughout a person's life based on shared culture, appearance, history and community values. From largest to the smallest, ethnic groups in Uganda include: the Buganda,[140] Basoga, Bunyoro, Batooro, Banyabindi, and Banyankole.

The above post is one example of the ethnicity-based hate speech we found. Local partners found it degraded and dehumanized the Buganda people, insulting their intelligence by comparing it to that of an animal.[141] Other ethnicities were also targeted. Additional terms from the lexicon characterize "bakonzo" into a hateful term by reinforcing negative stereotypes about the group. These adjectives in whole—rebel, maneater/cannibal, monkey, killer, fighter, wicked, "hard-bodied" (brutish), cruel to women, "short-minded" (unintelligent), and short—read like descriptions of the African savage depicted in colonial discourse. Used in conjunction with the neutral term Bakonzo, the adjectives are meant to vilify members as violent, animalistic and lacking in morality. However, the inclusion of "rebel" and "fighter" add an undeniable political and postcolonial dimension to the depiction.

139. *Id*.
140. Muganda is the singular form of Buganda.
141. Interview with Johncation Muhindo, *supra* note 76.

5. Classism

FIGURE 8. ECONOMIC GROUP

Kasese society creates a strong distinction between the haves and the have-nots. Notably, the economic divide dictates who has access to the internet and social media. Ugandan participants referred to others with education and economic prospects as "elites." Those who do not consider themselves upper-class often accuse elites of corruption. The above social media post originally discussed a corrupt businessman. In sharing this post, this user added language in the caption calling for the decapitation of the businessmen in the photo.

6 Ageism

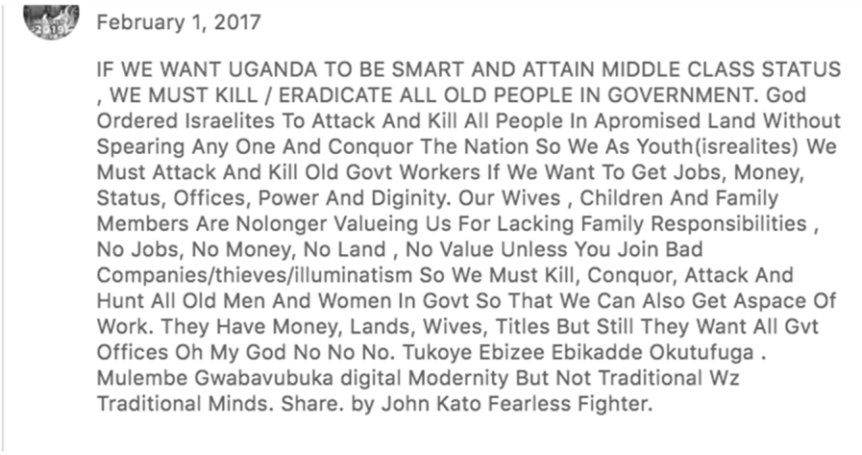

FIGURE 9. AGE SOCIAL MEDIA EXAMPLE

Age may be closely tied to elite status, as Uganda suffers a crisis of youth underemployment and unemployment. Beyond economic status, the above post also explicitly targets the elderly with violence. The poster calls to kill "old people in government" as a way to improve the Ugandan government. There is a long-running frustration with elderly politicians providing no political space for others. Political leaders are seen as entrenched; for example, President Museveni has been in power since 1986.[142]

7. Gender

FIGURE 10. GENDER SOCIAL MEDIA EXAMPLE

Participants in the Kasese community trainings cited gender-based discrimination as dominant in online forums. The first example of hate speech targets gender in the Kasese context. It sexually berates Bakonjo women. It

142. *See* ENCYCLOPEDIA BRITANNICA, *supra* note 73.

is also an example of content that may not be possible to categorize as hate speech without further localized knowledge on Kasese gender relations.

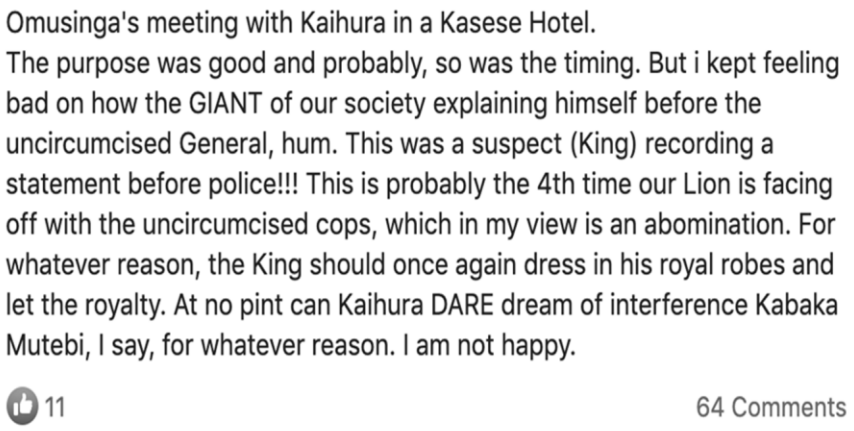

FIGURE 11. TRIBAL TRADITION AND GENDER SOCIAL MEDIA EXAMPLE

Tradition often refers to cultural practices specific to various ethnic or tribal populations. In Kasese, the most targeted tradition was circumcision. Among some populations, uncircumcised men are considered inferior. The Facebook post above exemplifies that attitude and distinguishes the usage from religious practices.

This example also demonstrates the intersectional nature of hate speech, as it combines tribal affiliation or tradition with gender. As will be discussed later, this is also a common characteristic found in postcolonial hate speech.

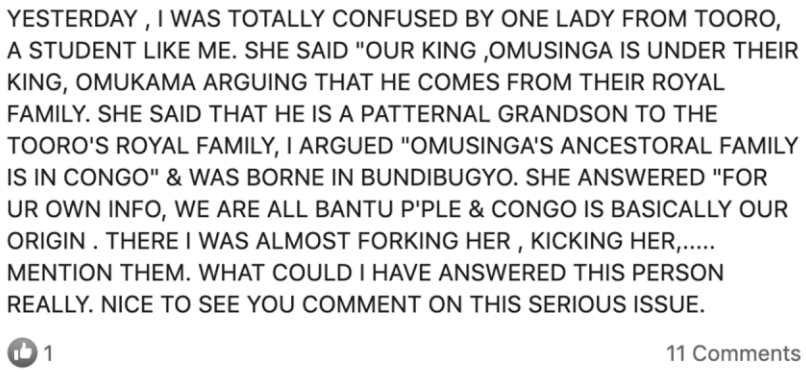

FIGURE 12. NATIONAL IDENTITY SOCIAL MEDIA EXAMPLE

Hate speech premised on national origin targets a person or group of people on the basis of belonging to a particular State or nationality. While none of the discovered content explicitly targets individuals based on national

origin, posts do reference this distinction implicitly, as indicated above. This content highlights the tensions between Bakonzo and Batooro tribes and the impact of their divisive history on present community relations. It also alludes directly to the colonial territorial origins, tribal affiliations, and shared linguistic Bantu roots in southern Uganda, and how contemporary borders do not align with traditional areas that groups occupied.

C. *Patterns of Hate Speech*

Distinct patterns in online speech and hate speech emerged in the course of the study and affected our evaluation of our methodology, our content collection, and our content analysis.

FIGURE 13. CODE-SWITCHING POSTINGS

English is the primary language of online discourse in Kasese. However, much of the content on social media exhibits a form of dialect, or code-switching between English and other Ugandan regional languages. Other posters of hate speech have been studied using this technique; for example, disgruntled users have attempted to evade AI content moderation systems, associating innocuous words or corporate brand names with hate speech-related meanings.[143]

Linguistically, parsing online speech is complex. The Kasese-based posts also exhibited diverse colloquial spellings. The mixed language communication complicated open-source research for the terms in the lexicon. For instance, the poster below wrote "ov da" instead of "of the."[144] Translation also

143. Fernando H. Calderon, Namrita Balani, Jherez Taylor, Melvyn Peignon, Yen-Hao Huang & Yi-Shin Chen, *Linguistic Patterns for Code Word Resilient Hate Speech Identification,* Sensors 2021 (2021).

144. Emmanuel Chukuwudi Eze, *Language and Time in Postcolonial Experience*, 39 RSCH. IN AFR. LITERATURES 24, (2008).

made it difficult to interpret content against the community's definition of hate speech. This challenge underscored the importance of working with individuals fluent in local contexts and languages when addressing hate speech online.

The prevalence of this content in hateful contexts may have been a deliberate strategy by the posters. Some of this content may have been "leetspeak," a strategy to use alternative spellings to evade keyword filters.[145] We did not find enough data, however, to ascertain whether the identified patterns were community-specific colloquialisms, efforts to evade content filters, or simple spelling errors.

D. *Importance of Comments*

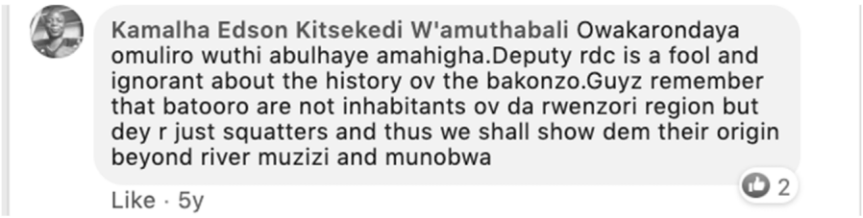

FIGURE 14. COMMENTS ON POSTS

Divisive or hateful content was often just as frequently or more frequently found in the comments rather than in the original posts.[146] Hate speech occurred especially in politically- oriented posts, as the corresponding comments were more likely to contain content that violated Facebook's community standards or implementation standards. Inflammatory political topics more often referenced local issues, like kingdoms and cultural institutions, tribal affiliations, and tribal royalty, than national electoral politics. This trend reflected the interrelated nature of political and tribal identities in Kasese. This could pose problems for content moderation strategies that seek to both protect freedom of expression and disincentivize hatred based on tribal affiliation, but do not understand the social positioning or role of tribes as distinct from ethnicity.

Searches within the comments also yielded significant quantities of counterspeech, or speech that attempted to mitigate hate speech, in the comments of posts. The counterspeech appeared in various forms. Some instances applied fact-centered rhetoric to defuse tensions and educate other discussants about broader issues relating to the negative speech. Some counterspeech made requests for more civil discourse. The following examples show how

145. Calderon et. al, *supra* note 143.

146. Comments are especially important as both Facebook's internal algorithms and content moderation algorithms are inaccessible to most external technologists and researchers.

content associated with hate speech also contained counterspeech. The first post states "But let's not abuse each other and learn from this post. If you have a belief that someone who isn't a Mukonjo (singular person of Bakonjo tribe) will lead the Bakonjo to the maximum—that's impossible. There will be repatriation of the resources to his homeland. Believe me or not, let's vote wisely."

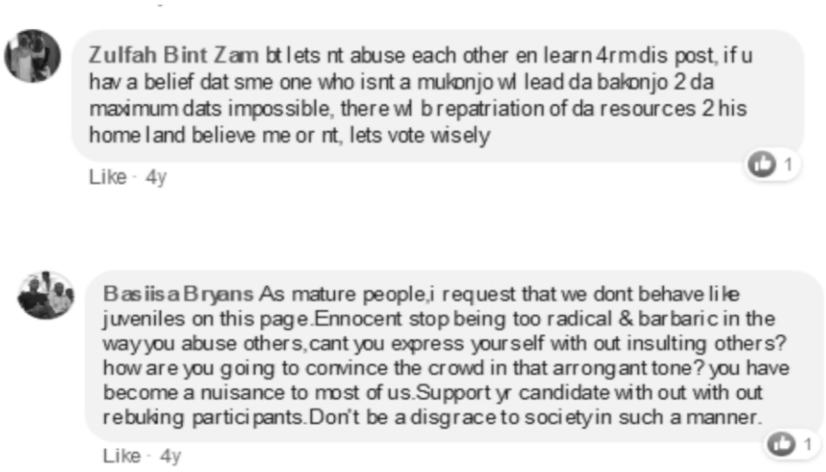

FIGURE 15. COUNTERSPEECH

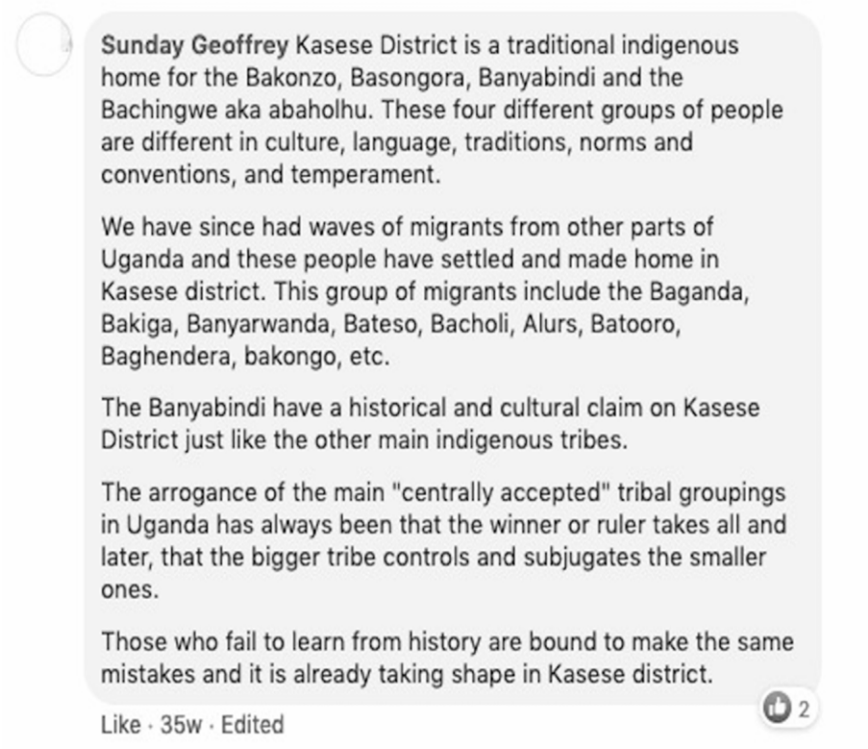

FIGURE 16. ADDITIONAL COUNTERSPEECH

IV. DISCUSSION

A. *Discussion: Online Hate Speech in a Postcolonial Context*

Where communities like those in Kasese continue to contend with the social, economic, and political legacies of colonialism, it is vital to interrogate specific features of hate speech that characterize how it manifests in a postcolonial context. Further research on hate speech in other postcolonial contexts is necessary to determine whether this connection to colonial legacies is widespread.

Because much of the theory around hate speech involves cultural stereotypes, some of the participants in the Kasese study expressed or implied concern that countering hate speech was tantamount to challenging traditional culture. Tensions evident in the hate speech in Kasese's online communities may trace back through generations of conflict, during which the stereotypes and negative attitudes of other local groups may have become deeply

ingrained. A workshop participant stated, "I don't want my daughter to grow up believing in culture. I want her to be empowered." This remark may reveal the participant's identification of an implicit contradiction between "culture" and "empowerment." This is reminiscent of past dynamics, when colonizers contrasted traditional practices with "progress" by conflating negative stereotypes with local culture at large.[147]

Presently, social media platforms designed mostly in the United States are used across the world in broadly diverse contexts. The platforms' design powerfully mediates communication and intimately affects the experience of individual users immersed in diverse cultures and political environments. Because platforms are optimized to maximize profit, social media companies prioritize growth and entry into new markets over concerns related to human rights and positive community engagement.[148] It is therefore vital to acknowledge the immense power that platforms wield in the design of their systems and interrogate the application of that power.[149]

Discussions with workshop participants revealed that platforms are associated with wealth, whiteness, globalization, and social progress, which have historically been held as the default or even subconscious locus of aspiration in poor communities around the world according to postcolonial theory.[150]

The practice of content moderation further complicates this fraught dynamic. Moderating content at the massive scale on which these platforms operate requires companies to create universally applicable governing policies. However, these policies may be ill-suited to diverse communities. Many social media platforms treat platform content as standalone—outside of time and devoid of history, rather than capable of interacting with other platforms' organic content or as part of a wider geopolitical communications and machinations.

As Benesch and many others have advocated, international human rights law provides a useful guide on which to base these policies.[151] As a basic

147. As an American woman facilitating the workshop and as an outsider to the community, I did not imagine this project as mediating long-standing social tensions. My local partners and I aimed to understand how tensions manifest on social media platforms and to explore the extent to which these platforms exacerbate those tensions. It is necessary, however, to acknowledge and contend with the colonial and racial dimensions of this work, and consider whether, as well-meaning researchers, we are inadvertently reinforcing these power dynamics. In future iterations, we hope to work toward a model in which local community leaders are wholly presenting material and outside researchers are simply facilitating the workshop, providing technical or legal knowledge as requested, and helping local leaders gain access to contact with platforms.

148. Paul M. Barnett, *Who Moderates the Social Media Giants: A Call to End Outsourcing,* N.Y.U. CTR. FOR BUS. & HUM. RTS. (June 2020), https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497.pdf.

149. Klonick, *supra* note 67 at 1601–03.

150. *See, e.g.*, FRANTZ FANON, A DYING COLONIALISM 71 (Haakon Chevalier trans. 1965) (1959).

151. *See* DANGEROUS SPEECH PROJECT, *supra* note 70 at 7–8.

framework, it has yet to gain a foothold in the creation and practical application of platforms' corporate policies. Because of regionally specific differences in international human rights law, it would be beneficial to include these instruments and priorities in platforms' calculus. Similarly, acknowledging how African versions of the sources and application of international law differ may help counter the influences of Eurocentric legal traditions and American platforms.

Furthermore, this inquiry may produce more questions around globalization, digital colonialism, and platforms' capitalist business models than it may answer.[152] But one thing is abundantly clear: Platforms alone do not have the answers to these questions.

Local perspective is indispensable in understanding hate speech. One of the greatest advantages of this study's methodology, a hybrid of local expertise and remote support, is the opportunity for the reciprocal exchange of knowledge and expertise between our Kasese-based partners and California-based researchers.

As previously described, our local partners were vital for gathering information and compiling our results. They offered deep insider knowledge of cultural stereotypes, lexicon items, the role of social media in communities, and local attitudes around hate speech. The project would not have been possible without these insights. Their input was also invaluable in parsing instances of content that exhibited code-switching between English and local languages and assisting with classifying content that was difficult to categorize.

Substantial research has also demonstrated that counterspeech is one of the most powerful tools for mitigating hate speech,[153] especially when platform-led moderation practices have limited efficacy. While the most impactful type of content counterspeech content is beyond the scope of this study, counterspeech education for partners was built into the curriculum in a module in the Kasese workshop.

Upon reflection, there were a number of limitations of the local-remote hybrid approach. In the two major phases of our workflow—content collection and content analysis—conducted by U.C. Berkeley-based researchers, there was substantial room for error in translating the community insight from our partners into content decisions. In future iterations of this methodology with greater resources for compensating partners, involving local partners in analyzing more (or even all of) the collected content would be more effective.

152. *See, e.g.*, Jacob Breslow, *Moderating the 'Worst of Humanity': Sexuality, Witnessing, and the Digital Life of Coloniality*, 5 PORN STUDIES 3 (June 6, 2018), https://www.tandfonline.com/doi/pdf/10.1080/23268743.2018.1472034.

153. DANGEROUS SPEECH PROJECT, *supra* note 70 at 25.

V. RECOMMENDATIONS

A. *Improvements to Content Moderation Structure and Staffing*

While the study focused on Facebook and the Kasese region, the problems that emerged may have wider applicability to different markets and other platforms, especially in the Global South.

1. Increase Localized Staff and Nonprofit Engagement

Global companies have a responsibility to understand the local context and the potential impacts of their platform in a community before they enter a market, begin to provide products and services, and profit from those users.[154] This level of local knowledge is only possible through close local engagement, which requires both a larger staff in the region and prioritizing data outflows from the market in times of greater risk, like around elections.[155] This may require investment in full-time staff and not just outsourcing more contractors. Some have even suggested creating a diplomatic-like structure with country directors and policy specialists where platforms operate. While the optimal structure of engagement has yet to be determined, local offices can better engage in meaningful partnerships with local nonprofits and other actors.[156]

Additionally, it is important to acknowledge the challenges that local offices present, to ensure they do not reflect the divisions that companies seek to prevent. There are issues of perception, if tribal or regional affiliation of an employee are seen to bias their judgment, and practical issues, like balancing groups in hiring decisions. While local knowledge and context is essential, companies should proceed carefully.

Though Facebook has over 2.4 million users in Uganda, the company does not have a locally informed presence in the country.[157] Content moderation efforts would be improved in local markets if Facebook established better employee representation where it operates, and further empowered those individuals to make policy changes, like adding to community standards. While employees are able to submit suggestions at biweekly meetings and add to the implementation standards, the community standards are the ultimate

154. U.N. OFFICE OF THE HIGH COMM'R FOR HUM. RTS., GUIDING PRINCIPLES ON BUS. AND HUM. RTS.: IMPLEMENTING THE UNITED NATIONS "PROTECT, RESPECT, AND REMEDY" FRAMEWORK, at 17–18, U.N. Doc. HR/Pub/11/04 (2011). https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf.

155. *See* Heller, *supra* note 8.

156. *See* Barnett, *supra* note 149 at 26.

157. *Id*. Facebook has not publicly confirmed where it has regional representation in Africa.

internal governance of content.[158] Tools like the U.N. Guiding Principles on Business and Human Rights can be applied to balance increasing local company presence against the risk of putting employees in country, especially in volatile locations where the government may use threats to these employees as a potential lever to pressure their employers.

2. Increase Transparency

Furthermore, Facebook would also increase public trust if it were more transparent about the distribution of its employees in volatile markets. With this study in mind, it would be useful to have publicly accessible answers to questions such as: How many salaried employees does Facebook have in Africa? In Uganda? How many are externally contracted content moderators? What are their respective roles in content moderation? How are they trained on hate speech identification and enforcement? What is the history of Facebook's partnerships with nonprofits and external researchers in the region? What are some examples of successful engagements?

Content moderation professionals—whether employees or contractors—unfamiliar with local regional context should participate in concrete community engagement for a sustained period of time. Specifically, moderators need to understand hate speech in the context of the relevant languages and culture from which the speech arose.[159] Additionally, those who design the community standards and implementation guidelines of platforms would benefit from engaging in active dialogues with local stakeholders, community members, and nonprofit organizations that understand the zeitgeist and cultural context of both online and offline speech.

B. *Acknowledge the Significance of Postcolonial Context*

Calls for geographic exclusion, as a new hallmark of dangerous speech, warrants further examination by platforms, specifically looking at the significance of colonial and postcolonial elements in hate speech emerging from this region. In countries like Uganda where Facebook is tantamount to the internet, this finding may warrant a reflection in the way that hate speech is categorized in the company's community standards and implementation standards. Internally, Facebook's Tier 2 of hate speech in the community

158. *Id.* at 1. According to Paul Barnett, "Facebook has arranged for additional outsourced moderators to pay attention to countries like Myanmar. Indonesia, and Ethiopia. This is a step in the right direction, and the expansion should continue until these countries have adequate coverage from moderators who know local languages and cultures— and function as full-time Facebook employees. Increased moderation needs to be accompanied by the presence of a country director and policy staff members in each country where Facebook operates. Responsible global companies have people on the ground where they do business. A social media platform should be no different. Facebook, YouTube, and Twitter should have offices in every country where users can access their sites." *Id.* at 25.

159. *See* discussion *supra* I.C.

standards prohibits calls for exclusion, so including geographic exclusion as another form of speech that is likely to incite violence could significantly aid Facebook's content moderation strategies.[160]

C. *Moderate Leetspeak and Colloquialisms*

As previously stated, our results included typographic irregularities that we initially believed could be some form of leetspeak.[161] But upon further contextualization, we hypothesized that these irregularities were likely more indicative of colloquial online speech patterns rather than intentional efforts to subvert algorithmic detection[162] of potential hate speech. Further research is necessary to confirm the validity of this hypothesis. Leetspeak deserves attention in content moderation efforts as a stylistic preference or an intentional effort to evade moderation algorithms. Platforms should conduct further research to determine whether these patterns are regularized or predictable, in individual cultural and linguistic contexts, so that they may be accounted for in algorithmic content moderation. Because leetspeak is a behavior rather than pure content, the importance of investigating leetspeak patterns underscores that in studying hate speech, platforms should examine actors and behaviors in addition to the content they generate.[163]

D. *Evaluate Relevant Legislation*

Platforms should consider creating nation-specific plans for content moderation that interact constructively with relevant national and local legislation. As previously mentioned, in Uganda, national legislation sponsored by President Museveni (under the guise of preventing the spread of disinformation) has resulted in suppression of speech through imposed internet blackouts and a social media tax to "control rumors."[164] Museveni even blocked Facebook, claiming the platform did not control disinformation in the following election. But what had really occurred was that Facebook removed a network of false

160. *Facebook Community Standards*, TRANSPARENCY CENTER, https://transparency.fb.com/policies/community-standards (last visited Feb. 3, 2022). https://transparency.fb.com/policies/community-standards.

161. Christian Espinosa, *Leetspeak: The History of Hacking Subculture's Native Tongue*, https://christianespinosa.com/blog/leetspeak-the-history-of-hacking-subcultures-native-tongue (last visited Jan. 3, 2023).

162. Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti & N. Asokan, *All You Need Is "Love": Evading Hate Speech Detection*, *in* PROC. OF THE 11TH ACM WORKSHOP ON A.I. SEC. (Jan. 15, 2018).

163. Camille François, Actors, Behaviors, Content: A Disinformation ABC 7 (Sept. 20, 2019) (unpublished manuscript) (https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf).

164. Carmel Rickard, *Uganda's Internet Closure During Elections Challenged at East African Court of Justice*, AFR. LEGAL INFO. INST. (Mar. 18, 2021), https://africanlii.org/article/20210318/ugandas-internet-closure-during-elections-challenged-east-african-court-justice; *see also* Nanfuka, *supra* note 103.

accounts designed to influence the election, stemming from Museveni's own government, under coordinated inauthentic behavior policies.[165] Understanding how capricious speech-oriented laws can be, claiming to target one social ill while being employed to limit expression and opposition, is vital for protecting international human rights. Museveni's type of tactics may evolve into access to information-related challenges as governments take increasingly aggressive stances against Facebook and other platforms. These legal dynamics influence patterns of speech and social dynamics, in both online and offline spaces.

E. *Increase Frequency of Human Rights Impact Assessments in At-Risk Regions*

Human rights impact assessments (HRIAs) are vital tools to understand the risks that a company's products or services bring to a specific market.[166] HRIAs began as audits to help industries like apparel, mining, and gas/petroleum extraction grasp the actual and potential harms to local populations caused by their operations. Technology companies began conducting HRIA with outside auditors in the early 1990s.[167] Now these impact assessments are increasing common, as organizations like the Global Network Initiative require their tech-company membership to undergo independent assessments, focused on freedom of expression, every two years.[168]

In places like Uganda, where there is a history of long-standing ethnic tensions and a recent history of ethnic and electoral violence, platforms should carefully conduct human rights impact assessments before harm is done. Additionally, the Global South deserves more attention due to the outsized weight of platforms like WhatsApp and Facebook in their societies.[169]

F. *Increase or Restore Resources for Open-Source Research*

From our project's inception to its implementation, Facebook limited researchers' ability to conduct open-source investigations. First, searches can only be text-based, excluding image-based searching that might have

165. Tessa Knight, *Social Media Disinformation Campaign Targets Ugandan Presidential Election*, DIGIT. FORENSIC RSCH. LAB (Jan. 11, 2021), https://medium.com/dfrlab/social-media-disinformation-campaign-targets-ugandan-presidential-election-b259dbbb1aa8.

166. FACEBOOK, *supra* note 11.

167. Michael A. Samway, *Business, Human Rights and the Internet: A Framework for Implementation, in* HUM. DIGNITY AND THE FUTURE OF GLOB. INSTS. 295, 309–12 (Mark P. Lagon & Anthony Clark Arend eds., 2014).

168. *Company Assessments*, GLOB. NETWORK INITIATIVE., https://globalnetworkinitiative.org/company-assessments (last visited Aug. 3, 2022).

169. *See* Mansoor Iqbal, *WhatsApp Revenue and Usage Statistics (2022)*, BUS. APPS, https://www.businessofapps.com/data/whatsapp-statistics (last updated Oct. 24, 2022); *see also* Mansoor Iqbal, *Facebook Revenue and Usage Statistics (2022)*, BUS. APPS, https://www.businessofapps.com/data/facebook-statistics (last updated Nov. 24, 2022).

uncovered valuable meme-related content. Second and more significantly, Facebook's API and interface proved to be limited during the discovery process because Facebook eliminated Graph Search[170] functionality in June 2019. This meant that the high degree of precision and accuracy within Facebook's search capabilities, including regional and temporal parameters, were no longer available. For example, temporal searches permit the tracking of trends around specific dates, such as election periods or instances of known offline violence. These searches are nearly impossible to complete without access to inside information from Facebook.

As a result of these technical challenges, our open-source research yielded smaller amounts of content than we likely would have encountered had we been able to further customize search strategies and content formats. We expect that application of our methodology to other social media platforms, such as Reddit and Twitter, which provide more customizable and accessible APIs, would likely yield a higher quantity and quality of publicly accessible content.[171]

Content moderation and the removal of harmful online content are necessary, but simply removing content without allowing further study by external researchers misses huge opportunities to better understand of this content. For example, this type of review could help researchers address the relationship between hate speech prevalence and platform design. Therefore, it would be beneficial for Facebook to provide vetted civil society researchers with increased access to anonymized content that has been taken down for violating community standards. This content has significant social and cultural value for researchers seeking to understand the nature, timing, and patterns of hate speech to mitigate its effects and develop better policies and practices

170. According to TechCrunch, Facebook Graph Search was a semantic search engine that was introduced by Facebook in March 2013. It was designed to give answers to user natural language queries, such as, "Who are my friends who live in San Francisco?" rather than a web-based search that would display a list of links. Drew Olanoff, Josh Constine, Colleen Taylor & Ingrid Lunden, *Facebook Announces Its Third Pillar "Graph Search" That Gives You Answers, Not Links Like Google*, TECHCRUNCH (Jan. 15, 2013), https://social.techcrunch.com/2013/01/15/facebook-announces-its-third-pillar-graph-search.

171. We chose to limit our scope of research based on various ethical considerations. We limited our scope of research by searching only in Public Facebook Groups and Pages. The U.C. Berkeley Human Rights Center holds a strict zero-interaction policy for all digital investigations to preserve the safety and security of student researchers. Though friending community members and entering closed groups would have yielded more robust data, it would have exposed student researchers to unnecessary security risks. Additionally, we determined as a team that entering Private Facebook Groups for our research purposes without disclosing those purposes would breach a barrier of consent that we felt ethically obligated to uphold. Facebook activity in Public Groups, whether implicitly or explicitly recognized, contain public discourse which anyone may see, access, and interpret. Private Groups, however, purposefully restrict access to those who are members of those groups. We expect that content found in Private Groups would have been more vitriolic and therefore quite useful to our research; nevertheless, we chose not to engage to preserve the ethical boundaries of our investigation. We found that at least in the Kasese context, there were a number of public Groups that served our purposes effectively.

for regulating it. Companies like Facebook can also improve partnerships with researchers and civil society by supporting the development of best practices, like enforceable ethical standards.[172] To do this most effectively, this paper proposes establishing and enforcing an ethical code for social media researchers, modeled off the Berkeley Protocol, to accommodate the global nature of hate speech and the needs of those most impacted.

As previously mentioned, search functionality on Facebook has been extremely limiting for open-source academic researchers. To improve social scientists' ability to understand how platforms function, Facebook should improve time-bound searching. This would make it possible to examine temporal shifts in language and speech surrounding particular events, enabling tracking of the trajectory of influence from online speech to offline action. Facebook should also index and create search capability of text in images. This functionality would allow better categorization and analysis of image-based speech such as memes.[173]

G. *Connect Hate Speech and Other Types of Platform Enforcement*

Platforms would also benefit from creating policies to enforce and dissuade state-sponsored hate speech. At the time of this writing, Facebook remains blocked in Uganda. After the platform conducted a takedown of a network pages and users tied to the Ugandan Ministry of Information, the government blocked Facebook during the electoral period. However, a substantial time later, the platform has not yet been reactivated within Ugandan borders.

Setting clear enforceable expectations about governments misusing and manipulating online social media networks would create a safer environment for users, but may cross over into policies in other areas, like inauthentic behavior, and areas sorely needing policies, like state-sponsored harassment.[174] As previously noted, this may result in a different tenor of responses, as hostile governments have increasingly sought to penalize social media companies and their employees for content enforcement. This type of government targeting of civilians could be mitigated by more direct company policies setting out consequences for inauthentic behavior.[175] However, focusing on state-sponsored harassment still risks a deflecting response like Museveni's—blocking

172. An example of this type of guidance is the Berkeley Protocol on Digital Open Source Investigations (2020). However, the content of such a new protocol is outside the scope of this paper's inquiry.

173. *See* Douwe Kiela, Hamed Firooz & Aravind Mohan, *Hateful Memes Challenge and dataset for research on harmful multimodal content*, META AI (May 12, 2020), https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set.

174. *See* Brittan Heller, *Enlisting Useful Idiots: The Ties Between Online Harassment and Disinformation*, 19 COLO. TECH. L.J. 19 (2021).

175. *See generally Inauthentic Behavior*, META, https://www.facebook.com/communitystandards/inauthentic_behavior (last visited Jan. 4, 2023).

access to information based on an alleged social media enforcement action, which is clearly a screen to justify his government's actions. While more can always be done to increase clarity, additional study needs to be done to determine what may actually deter this type of government response.

VI. CONCLUSION

Hate speech moderation cannot be done effectively in the abstract. This project offered a unique opportunity to gain a deep understanding of hate speech in a specific locality by engaging in on-the-ground fieldwork. Conducting a workshop and engaging directly with those most impacted by hateful speech, with native knowledge of local languages and culture, provided insights that would have been impossible to derive from remote open-source research alone.

Through the open-source phase of this project, we built a database of 100+ potential instances of hate speech from public Facebook groups linked to Kasese and identified general patterns and tendencies of hate speech in the Kasese context. However, the limited duration of the time in Uganda, limited extent of the contextual research, and team members' backgrounds made it impossible to develop a comprehensive understanding of the context sufficient to definitively categorize all potential instances of hate speech. It underscored the importance of having platforms support this type of research and use the results to develop durable yet flexible content moderation policies.

This project confirmed the indispensable nature of localized perspectives when engaging in content moderation, or in research investigating the impact of digital technologies on local communities more generally. It also highlighted the necessity of nuance: content cannot be meaningfully categorized as hate speech without social, political, cultural, and historical knowledge.

Kasese remains just one region out of many others across the globe where content moderation strategies fall short of what is effective to deescalate hateful online discourse. Community members in Kasese who are already educating each other and spreading counterspeech are meaningful examples of what social media companies could do to address hate speech on their platforms – integrate local and global priorities, especially in postcolonial contexts.

APPENDIX I: SELECTED LEXICON OF HATE SPEECH IN KASESE

By asking teams of participants to brainstorm slurs and negative stereotypes against the groups listed in the community definition, U.C. Berkeley's Human Rights Center and I generated a vetted list of forty-seven words and phrases in Lukhonzo, Lutooro, Kiswahili, and English related to potential hate speech. Below are ten examples.

Example Term	Definition
Ebisiyaga	Derogatory term for homosexuals
Omutsule, Abat[s]ule	Uncircumcised (singular/plural)
Abanyagwagwa	Foreigners
Abalihanda	Used to refer to outsiders, other tribes
Rebel, man-eater, killer, wicked, fighter, hard-bodied + Bakonzo	Stereotype that Bakonzo are rebels, man-eaters, monkeys, killers, wicked, fighters, hard-bodied, and cruel to women
short + Bakonzo	Stereotype that Bakonzo are short and short-minded
Ekithaka Nikyethu	"Land is ours"; expresses territorialism
Emalaya	Prostitute
Endaghangali	Traitor: in Musoki Elizabeth's words, during the Rwenzururu struggle from 1962–1982, "if you were called such it meant you betrayed the cause and sided with the enemy against your own."
Omulihanda	Enemy, usually of the enemy tribe