The Legal Journal on Technology
Public Comment: Case 2025-044-FB-UA, 2025-045-FB-UA, 2025-046-FB-UA, 2025-047-FB-UA (Content Moderation in Somaliland Cases)

Summary: In the "Reporting on Somaliland Current Affairs" case before Meta's Oversight Board, The Legal Journal on Technology (TJLT) argues that protecting media freedom in non-English speaking and politically sensitive regions requires systemic reform to prevent unjust censorship of journalistic content. As outlined in TJLT's response, the landscape in Somaliland is marked by the persistent suppression of independent media, despite constitutional promises of free expression. Journalists face intimidation, legal pressure, and the constant threat of digital repression, particularly on global platforms. TJLT contends that Meta's automated moderation tools, predominantly trained on Western datasets, routinely misclassify critical journalism in low-resource languages like Somali.

This not only silences vital watchdog voices but also perpetuates structural marginalization. TJLT identifies the need for procedural fairness, transparency, and culturally competent content moderation. The submission advocates for concrete safeguards such as including mandatory human review for public-interest news pages and a novel "Public-Interest Media Flag" to protect high-risk journalists in crisis zones. It recommends clear notification protocols, specialized appeal pathways, and transparency measures to restore trust when errors occur. Ultimately, TJLT asserts that rapid, tailored remedies and preventative mechanisms are essential to uphold journalistic freedom and due-process rights in digital spaces, ensuring media pluralism is preserved where it is needed most.

ISSUE 1: MEDIA FREEDOM AND SAFETY OF JOURNALISTS IN SOMALILAND, THE ROLE OF SOCIAL MEDIA AND THE SITUATION FOR FREEDOM OF EXPRESSION.

TJLT Response: Somaliland is a self-declared republic that broke away from Somalia in 1991, which currently operates independently with de facto autonomy. Freedom of independent media and journalist safety continue to be long-awaited ideals and dreams that are yet to be realized. While Somaliland citizens have been formally promised freedom of expression and press independence in the Constitution, these protections are still only on paper.[1] On the ground, journalists face significant barriers in operating independently and disseminating information. Over the past decade, the Somaliland government has adopted an increasingly restrictive stance toward independent media, particularly in electronic platforms. Private radio stations have effectively been banned, based on the government's claim that open broadcasting would be disastrous in Somaliland's "highly argumentative society." This rationale, unfortunately, has been used to justify widespread suppression. One prominent example of this is Radio Horyaal, a privately-run station that was forced to relocate to Belgium to continue operating, following repeated harassment and denial of registration in Somaliland.[2]

Broadcasting via television does not fare well either. Although television broadcasters like Universal TV and Horn Cable TV are technically permitted to operate, they function under constant state scrutiny. They are routinely accused of bias and subjected to political pressure. The broader pattern is clear: journalists and media platforms that do not align with the narrative set by the state are met with interference,

[1] Constitution of the Republic of Somaliland 2000, art 32.
[2] Nicole Stremlau, 'Hostages of peace: the politics of radio liberalization in Somaliland' (*Somaliland Economic*, 13 May 2025) <https://somalilandeconomic.com/hostages-of-peace-the-politics-of-radio-liberalization-in-somaliland/> accessed 9 July 2025.

intimidation, or legal consequences intended to hold control over information. These restrictions are maintained through aggressive tactics. According to the Somali Journalists Syndicate, at least 18 journalists were arrested in 2019 alone, many in incidents involving excessive force.[3] There are documented cases of police firing on journalists, seizing equipment, and deleting footage of politically sensitive events. Such practices have created a climate of fear and self-censorship among journalists.

The judiciary further aggravates these challenges. Many judges frequently uphold convictions under vague or extremely broad charges like 'insulting public officials' or 'spreading false news.' Sentences in such cases range from several months to multiple years, further eroding confidence in legal protections for journalistic work. Social media platforms, especially Facebook, have emerged as vital tools for expression; however, they have simultaneously become dangerous spaces of repression as well. Journalists have faced imprisonment for critical posts addressing government policies, officials, and even public events.[4]

This digital repression goes beyond individual content. News websites have been blocked without a formal process, and digital reporters are often arrested for the material they publish online. In some cases, state authorities have relied on traditional clan structures to suppress dissent - by pressuring elders to serve as guarantors of a journalist's future conduct.[5] This practice by Somaliland exploits community norms to bypass legal accountability and silence critical voices.

For global platforms operating in regions like Somaliland, including Meta, it is essential to engage with local dynamics in a nuanced and informed manner. Wrongful content enforcement, especially against local journalism in politically sensitive and non-English-speaking contexts, risks reinforcing the very forms of repression it aims to avoid.

ISSUE 2. CHALLENGES IN PREVENTING WRONGFUL ENFORCEMENT AGAINST JOURNALISTIC CONTENT IN NON-ENGLISH CONTEXTS

TJLT Response: The submission further argues that freedom of press and media in non-English speaking regions faces risks due to wrongful enforcement. Automated content moderation systems, which operate on Artificial Intelligence whose data sets are often western-centric, limiting their ability to understand context, language nuance and journalistic expressions from other languages and cultures.[6] This leads to frequent misclassification of legitimate reporting as harmful or illegal content, resulting in the takedown of journalists' posts and accounts.[7]

In regions where freedom of press is restricted, these moderation errors assume greater significance. In a place where a journalist is less likely to cover a politically sensitive topic due to authoritarian governments, the recourses are also limited against content moderation by Meta. Under-resourced content moderation teams, often unequipped for determining the validity of texts from other languages other than English,

[3] Mohammed Dhaysane, 'Somalia: Attacks against journalists continued in 2019' Anadolu Agency (Ankara, 3 Jan 2020) <https://www.aa.com.tr/en/africa/somalia-attacks-against-journalists-continued-in-2019/1690701> accessed 9th July 2025.
[4] Reporters Without Borders, 'Three journalists held incommunicado and tortured in Somaliland' RSF (Paris, 11 January 2024) <https://rsf.org/en/three-journalists-held-incommunicado-and-tortured-somaliland> accessed 9 July 2025.
[5] National Union of Somali Journalists, Report on Somaliland: Media Freedom Kept within Bounds (2009) <https://rsf.org/sites/default/files/NUSOJ_Somaliland_Report.pdf> accessed 9 July 2025.
[6] Kinga Pázmándi, 'Theory and Practice of Social Media's Content Moderation by Artificial Intelligence in Light of European Union's AI Act and Digital Services Act' (2025) *EJ Politics* <https://www.ej-politics.org/index.php/politics/article/view/165> accessed 15 July 2025.
[7] Akshaya J and Nithya Sambasivan, 'Lost in Translation: How Content Moderation Fails Tamil Speakers Online' (2025) *arXiv preprint* <https://arxiv.org/html/2501.13836v1#S1> accessed 15 July 2025.

frequently confuse between hate speech and regional journalism, silencing critical and marginal voices.[8] In these scenarios, journalists may often find themselves locked out of their audiences, with little recourse and lack of timely redressal even if the issue is recognised. This vulnerability is amplified by historical power imbalances and the continued marginalization of non-English languages within platform policies.

Effective solutions require investment in linguistic and cultural expertise, transparent appeal mechanisms, and multi-stakeholder collaboration to ensure moderators, both human and automated. Without corrective action, wrongful enforcement will continue to threaten media freedom and pluralism, particularly where journalism is already under siege.

ISSUE 3. GOOD PRACTICES FOR ENSURING ACCESS TO ADEQUATE REMEDIES FOR JOURNALISTS AND MEDIA ORGANIZATIONS LOCKED OUT OF PAGES OR ACCOUNTS AS A RESULT OF WRONGFUL ENFORCEMENT.

TJLT Response: In the present case, Meta restored the content, acknowledging an error.[9] What remains unresolved is the structural question: *How should platforms respond to wrongful enforcement, ensuring journalists can access effective remedies?*

The posts reported on official events in Somalia, a politically sensitive and contested region.[10] In such a context, reporting on state activity often carries political or symbolic significance. However, as Meta later confirmed, contained no hate speech or incitement and were removed in error. The user stated clearly that their intention was to share public information, not to discriminate or provoke.

The implications of enforcement in such cases go beyond temporary account disruption. When a journalist loses access to their page and audience, it disrupts their ability to report on political developments, particularly in regions where traditional media is constrained. Remedy must therefore reflect not only a technical correction but the public role of journalistic speech and the disproportionate risks involved when removed without cause.

Meta's Community Standard on Hateful Conduct prohibits attacks based on protected characteristics, such as ethnicity and national origin.[11] However, it also recognises that not all references to identity are inherently hateful. Where journalistic content engages with political or social themes, enforcement must consider the context and the intention of the speaker.[12] This shows a disjuncture between the policy's intention and its application. Where enforcement results in wrongful unpublishing of journalistic pages, good practice should focus on procedural fairness, transparency, and institutional learning. To that end, several practices are worth adopting.

Meta's Oversight Board has identified *treating users fairly* and *refining automated enforcement* as strategic priorities.[13] In line with this, the first recommendation is to ensure mandatory human review of any

[8] Amrutha Mohan and V Padmaja, 'Lost in Translation: How Content Moderation Fails Tamil Speakers Online' (*Tech Policy Press*, 10 January 2024) <https://www.techpolicy.press/lost-in-translation-how-content-moderation-fails-tamil-speakers-online/> accessed 15 July 2025.
[9] 'Reporting on Somaliland Current Affairs' (*Oversight Board*, 1 July 2025) <www.oversightboard.com/pc/reporting-on-somaliland-current-affairs/> accessed 15 July 2025.
[10] Reporters Without Borders, 'Somalia' (RSF, 2024) <https://rsf.org/en/country/somalia> accessed 15 July 2025.
[11] 'Community Standards: Hateful Conduct' (*Meta Transparency Center*, May 2023) <https://transparency.meta.com/en-gb/policies/community-standards/hateful-conduct/> accessed 15 July 2025.
[12] UN Human Rights Committee, 'General Comment No 34: Article 19: Freedoms of Opinion and Expression' (12 September 2011) UN Doc CCPR/C/GC/34, paras 42 and 46 <www.refworld.org/legal/general/hrc/2011/en/83764> accessed 15 July 2025.
[13] 'Oversight Board Announces Seven Strategic Priorities' (*Oversight Board*, 13 March 2024) <www.oversightboard.com/news/543066014298093-oversight-board-announces-seven-strategic-priorities/> accessed 15 July

enforcement targeting media actors before removing pages that regularly post public-interest journalism. This can be triggered by posting patterns, regardless of verification status. Such pages should be routed into a *specialised review path* that includes contextual analysis by trained policy staff.

Second, the affected journalists should receive a policy-specific notice that clearly cites the Community Standard applied, identifies whether enforcement was triggered by user reports or automated systems, and outlines what kind of review (human or automated) will follow. Additionally, Meta should operationalise a *"fast-track review lane"* for accounts that

(a) consistently report on political events, and
(b) have no prior violations.

When such accounts are locked, they should automatically be routed to human moderators with training in public-interest content review.

Third, *a log of reversed media-related enforcement decisions* should be maintained internally and audited regularly to improve reviewer training and automated enforcement thresholds. In regions with heightened political sensitivity, Meta can utilize *local advisory input* from human rights groups and digital journalism networks to *improve classifications of political reporting*.

Finally, when a media page is unpublished and later reinstated, restorative remedy should not stop at access. Journalists should be offered the option to issue a reinstatement notice that explicitly explains the review process and affirms the content's compliance with platform standards. This feature wouldn't serve as an apology but rather as a *transparency measure*, helping journalists clarify to their followers that their reporting remains accurate and compliant, even after an inadvertent removal. Such clarity helps reduce reputational damage and restore trust.

Ensuring that journalists and media organizations affected by wrongful enforcement receive proper remedies, beyond simply restoring access, requires thoughtful, transparent, and respectful practices. Looking forward, strengthening remedies for journalists on social media is essential to uphold freedom of expression and press, ensuring digital spaces remain safe, credible forums for truth, accountability, and public voices.

RECOMMENDATIONS

Somaliland's media ecosystem is already under siege: at least forty-one Somali-language reporters have been arrested, intimidated, or assaulted since mid-March 2025, most for coverage of security or foreign-policy dossiers.[14] When Meta mistakenly "unpublishes" a 90,000-follower news page, it extinguishes a watchdog voice for roughly 1 in 70 citizens of the six-million-strong polity.[15] The Oversight Board's own cross-check opinion underscores that current mistake-prevention privileges politicians and celebrities, while local newsrooms remain unprotected. Below are two *lowest-friction*, yet world-class safeguards that can be deployed inside one product cycle to curb such collateral damage.

1. Lightweight Public-Interest Media Flag (PIMF)

2025.

[14] Committee to Protect Journalists, 'Alarming escalation: At least 41 journalists targeted since March in Somalia' (*CPJ*, 15 May 2025) <https://cpj.org/2025/05/alarming-escalation-at-least-41-journalists-targeted-since-march-in-somalia/> accessed 15 July 2025.

[15] M.A. Egge, 'Somaliland's population reaches 6.2 million' (*Horndiplomat*, 19 April 2024) <www.horndiplomat.com/2024/04/somalilands-population-reaches-6-2-million/> accessed 15 July 2025.

Meta already parses page metadata and runs language-ID scores for integrity workflows. A simple heuristic can elevate high-risk journalism without a line of new machine-learning code. A page is auto-labelled PIMF when it (a) self-identifies with "journalist/news/media" in its bio or domain e-mail, and (b) has ≥ 50,000 followers and ≥ 60 % of recent posts in a low-resource news language, such as Somali. Any proposed "unpublish" for a PIMF page is automatically paused until a *second* Somali-speaking reviewer confirms the violation. Because the rule is declarative, policy engineers can ship it via an Integrity rule-file update, avoiding the protracted model-training cycles that often stall Global-South fixes. The Santa Clara Principles demand upstream due-process guardrails; PIMF supplies them with almost no engineering lift.[16]

2. Twelve-Hour Soft-Stop with Shadow Appeal

When a PIMF post is suspected of breaching policy, Meta should apply a soft-stop rather than an immediate takedown. The content stays live behind a click-through interstitial and its reach is throttled by 75%, while a twelve-hour countdown notifies both editors and reviewers. Editors see a prominently placed *one-tap shadow-appeal* button that forwards the case to a priority queue without extending the deadline. If no definitive decision emerges within twelve hours, the post automatically regains full distribution and the strike is voided. This mirrors proven "hold-and-verify" patterns from financial fraud prevention, minimizing public-interest disruption while still containing potential harm, and squarely fulfils Access Now's call for accelerated human-review pathways in crisis regions.[17]

Together, PIMF functions as a proactive brake, while the soft-stop/appeal protocol operates as a reactive cushion. They translate the Santa Clara trilogy- *numbers, notice, appeals*- into safeguards tailored for low-resource languages, without the overhead of new neural architectures or extensive policy rewrites. Implementing these two layers would convert ad-hoc reversals into a rights-preserving workflow, strengthening Meta's compliance with international due process norms and materially improving the safety of frontline journalists in one of the world's most information-starved regions.

[16] 'Santa Clara Principles on Transparency and Accountability in Content Moderation' (*Santa Clara Principles*, 2022) <https://santaclaraprinciples.org/> accessed 15 July 2025.
[17] Access Now, 'New: Content Governance in Crises Declaration' (*Access Now*, 20 February 2024) <www.accessnow.org/publication/new-content-governance-in-crises-declaration/> accessed 15 July 2025.