WORK AND DIGITALISATION 25.11.2025 | Zikora Ibeh

# No sweet rape

## Algorithms meant to protect users now silence those who speak plainly about sexual violence, forcing euphemisms where truth should stand

Social networks have become defining features of public life. What emerged as casual spaces for entertainment and connection have matured into contested terrains where people claim authorship of their lives. Inside these digital commons, communities long denied legitimacy are carving out the fragile right to speak what repressive societies refused to listen to. But what happens when the very platforms that once appeared to widen the field of expression begin to reproduce the same architecture of restrictions that govern the offline world?

It was this contradiction that compelled Jude Eze, a cinematographer, to ask on Facebook, 'Why are we replacing the word "rape" with "grape"?' His question captured a growing trend. 'It is an ugly word for an even uglier act', he wrote. 'But that is precisely why it should be spoken in its full severity so that society is forced to confront the depth of its cruelty and the inhumanity it names.' Immediately, his comment section erupted into a chorus of shared frustration from users who had watched their posts flagged, shadow-banned or deleted — not because they endorsed violence but because they named it clearly and inadvertently crossed the platform's tripwires of content deemed sensitive or harmful. Many, myself included, recalled receiving stern automated warnings that our accounts were at risk for posting testimonies of sexual or gender-based violence. This phenomenon, which has been confirmed as a global digital experience, has enforced a linguistic contortion among social media users.

Rather than risk blowback, users, especially women already navigating precarity, now intentionally misspell rape, fracture it with asterisks, substitute the description with clusters of grapes or purple emojis, or erase entire accounts of violence altogether for fear of losing cherished pages that have become vital extensions of their lives. This evasive vocabulary belongs to a widening lexicon known as algospeak, a set of coded language adopted by users to avoid getting blocked. For instance, terms like 'kill' are now expressed as 'unalive', 'sexual assault' as 'SA', and 'suicide' as 'sewerslide'. And just like rape, these words are sometimes

further altered by creative spellings or representations, depending on how users think they can best slip past the system. But how did we get here?

## Machine eyes, human costs

The policing of content did not materialise out of nowhere. It grew out of a crisis that marked the earliest years of social media, when unregulated digital spaces became fertile ground for graphic violence, hate campaigns, extremist recruitment, and coordinated bullying and harassment. Governments, civil society actors, and other stakeholders warned that platforms left entirely to themselves could amplify real-world harm, and public pressure mounted for stronger oversight. It was within this climate that companies like Meta, which owns Facebook and Instagram, embarked on building systems capable of moderating harmful words or content and outrightly removing them where necessary.

The challenge was always volume. Millions of posts surfaced every minute, far beyond the capacity of the existing human workforce to evaluate, leaving moderators roughly 10 seconds to make a judgment call on each piece. Faced with this impossible pace, companies increasingly turned to automated systems that promised the speed and scalability needed to keep up.

What is now known as social media algorithms when it comes to content moderation are essentially sets of instructions that allow computers to identify patterns and make predictions. They are trained on large datasets to detect indicators of harm and operate through classification, a process that sorts content into categories such as safe or unsafe. They are powerful and quick, yet unable to grasp nuance or intention. And this is where the problem begins. The result is a machine logic that collapses radically different types of speech into a flattened zone of suspicion. Even Meta conceded that its automated systems often 'get things wrong' because they cannot interpret context in the way human beings do, and that its appeal processes are 'frustratingly slow' and 'do not always get to the right outcome'. In response to this problem, the tech giant is now shifting to a 'Community Notes' model — an approach also used by X that transfers some responsibility to users by allowing them to annotate content.

*The erosion of human oversight consolidates the authority of systems whose interpretive imagination is*

Yet the chronic failures of algorithmic judgement cannot be separated from the broader political economy of digital monopolies. Big Tech has spent years thinning out the very workforce that could correct these errors, instead of strengthening it. Across

*limited, intensifying the risks for communities whose survival depends on context being understood rather than flattened.*

different platforms, successive waves of layoffs reveal an industry intent on offloading its most delicate interpretative responsibilities to machines while recasting the humans capable of reading complexity as disposable. TikTok provides a recent example. Earlier this year, the company laid off hundreds of moderators, framing the cuts as part of a transition to AI-driven content filtering. Workers were told they were being replaced for efficiency even though these same automated systems routinely misclassify trauma testimonies, political satire, minority language expressions and basic news reporting.

These dynamics matter in the overall analysis because the erosion of human oversight consolidates the authority of systems whose interpretive imagination is limited, intensifying the risks for communities whose survival depends on context being understood rather than flattened. Nowhere are the costs of these limitations and expendable logics sharper than in Africa and other environments marked by entrenched gender inequalities and social volatility. The trend I now refer to as the 'fruitilisation of rape', which manifests online as the translation of unspeakable crime with sweetened symbols, is one of the clearest illustrations of how restrictive automated actions can become instruments of harm. This linguistic distortion trivialises violence and intensifies the emotional weight carried by those already struggling to be heard, especially in contexts where rape is chronically underreported due to stigma, threats and the slow churn of judicial systems.

Worse still, misogynists and digital misfits who understand that euphemised vocabulary provides cover for harmful rhetoric now exploit this vacuum, bandying the grape metaphor as jokes or threats and reproducing violence in spaces where survivors are penalised for naming what happened. If this is the reality, then the conclusion must be honest. There can be *no sweet rape*, and the task is not to invent gentler metaphors but to dismantle the architecture that demands them.

## What must change

The transformation required begins with a materialist reorientation of what platforms understand as their core obligation. Without a structural reset that centres user safety, narrative dignity and democratic accountability, any reform will remain cosmetic. Repairing the harm also demands restoring human judgment as the backbone of moderation. It

means retaining a robust global workforce of trained, adequately compensated and psychologically supported humans who can exercise careful judgment without the tyranny of impossible quotas. Machines may assist the process, but they cannot be entrusted with moral authority. Strengthening labour protections sits at the heart of this work, too, especially for outsourced moderators in the Global South whose rights are steadily eroded by layers of exploitative subcontracting practices.

Equally important is redistributing editorial power. The design of AI moderation cannot be based on datasets drawn from the Global North alone. Platforms must mainstream the perspectives of user collectives, survivors, rights advocates and other affected communities into the processes that shape how content is judged. Their expertise should inform the design of datasets, development of context guidelines and calibration of algorithmic thresholds. A humane digital environment must also enshrine protected testimony as a right. Users should be able to tag their content in a way that signals clearly that they are naming violence, not perpetrating it. Such an opt-in preserves clarity of language against literal-minded deletions that erase trauma.

As platforms turn to community-driven moderation systems, the shift carries real promise because it restores a measure of interpretive authority to end-users. But without strong safeguards, it can create new problems. Community judgment arising from societies steeped in patriarchy, homophobia, class prejudice or moral stigma can easily replicate the very injuries survivors already endure. Any transition to community-led oversight must therefore be transparent, globally representative and reinforced by strong safeguards that prevent hostile majorities from deciding whose testimonies are recognised as legitimate. Only by refusing euphemistic refuge and insisting on naming violence clearly can we reclaim the civic digital space as an environment where truth is not punished but protected.

---



## Zikora Ibeh

Zikora Ibeh is a researcher, columnist, podcaster and development advocate with a passion for social justice and gender equity. She works to make a difference in society through public policy advocacy, action research and media advocacy.