Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.17, No.1, 2026

Examining the Limitations of Existing Content Moderation Systems in Detecting Sheng-Based Cyberbullying on Kenyan TikTok

Kibet Dennis Kangogo[1*], David Muriuki Gikunju[2] David Mbogori[3]

1. School of Computing and Mathematics, The Cooperative University of Kenya, 24814 – 00502, Karen, Nairobi, Kenya
2. School of Computing and Mathematics, The Cooperative University of Kenya, 24814 – 00502, Karen, Nairobi, Kenya
3. School of Computing and Mathematics, The Cooperative University of Kenya, 24814 – 00502, Karen, Nairobi, Kenya

* E-mail of the corresponding author: mailto:kangogo.dennis@student.cuk.ac.ke

Abstract

TikTok has become the leading social media platform among young Kenyans, but its growth has come with a rise in cyberbullying, much of it written in Sheng or in comments that mix English, Swahili, and Sheng within a single sentence. The moderation tools used to filter harmful content are mostly trained on English data drawn from Western platforms, and earlier research suggests they miss as much as 68 percent of cyberbullying written in this style. This paper examines the scale of that failure directly, by testing four widely used content moderation tools, Perspective API, TextBlob, a generic BERT classifier, and RoBERTa, on a set of 5,000 Kenyan TikTok comments labelled by trained Kenyan linguists. The results show that all four tools fall well short of acceptable performance, with accuracy ranging from 44.7 to 64.8 percent and false negative rates between 43 and 67 percent. A culturally aware alternative built on the AfriBERTa architecture, included in the comparison for reference, reached 93 percent accuracy and an 8 percent false negative rate, underlining the scale of the gap. The findings confirm that current commercial moderation systems are not equipped to protect Kenyan TikTok users, particularly those communicating in Sheng, and that locally calibrated alternatives are both necessary and achievable.

Keywords: content moderation, cyberbullying, Sheng, code-switching, TikTok, natural language processing

DOI: 10.7176/CEIS/17-1-07
Publication date: August 28th, 2026

1.0 Introduction

TikTok is now the most widely used social media platform among Kenyan youth, with the Communications Authority of Kenya reporting over 12 million active Kenyan users and 85 percent internet usage among urban youth. Alongside this growth, cyberbullying has become more common, with the Kenya Institute of Counselling and Psychological Services recording an increase in cyberbullying rates among teenagers from 28 percent in 2020 to 42 percent in 2023.

Much of this harm passes through undetected because the systems built to moderate content are designed around English and trained mostly on data from Western platforms. Kenyan TikTok users frequently switch between English, Swahili, and Sheng, an urban creole spoken mainly by youth in Nairobi, sometimes within a single comment. Existing moderation tools generally struggle to interpret this mixture. The problem is made worse by how quickly Sheng changes, with roughly 40 percent of abusive terms shifting in meaning or spelling within six months, which leaves static keyword-based filters quickly out of date.

This paper addresses the first objective of a wider study into Kenyan cyberbullying detection: establishing, through direct empirical testing, how badly existing content moderation tools perform on Sheng-based and code-switched Kenyan TikTok comments. Understanding the scale and pattern of this failure is the necessary starting point for any effort to build a better alternative, since it identifies precisely where current systems break down and why.

2.0 Literature Review

This review moves from global findings to the African continent, then East Africa, and finally Kenya, in order to place the local situation within the wider body of research on content moderation.

2.1 Global Context

Commercial moderation systems generally rely on keyword filters and machine learning classifiers trained on large English datasets, most often drawn from Twitter and Reddit. Waseem, Thorne, and Bontcheva (2024) reviewed 87 studies on hate speech and cyberbullying detection and found that this reliance on English, Western-centred data limits how well such systems generalise to other languages and cultural settings. Fortuna, Nunes, and Soler-Company (2024) found that false negative rates exceed 40 percent outside English-speaking contexts, largely because these systems cannot read sarcasm, irony, or culturally specific phrasing. Kowalski and colleagues (2024) add that moderation systems are especially weak where young users employ subcultural or ironic language that looks harmless on the surface.

2.2 African Context

The same weaknesses appear more sharply across Africa. Ogunremi and colleagues (2024) tested five commercial moderation tools on 20,000 comments drawn from Nigerian, Ghanaian, and South African platforms and recorded failure rates between 55 and 72 percent for content written in local vernaculars, pidgins, and code-switched text. Adewole, Babatunde, and Okonkwo (2024) documented a similar pattern for Yoruba and Pidgin English on Nigerian platforms, describing the deployment of Western moderation tools in African settings as a form of algorithmic cultural exclusion, where the language patterns of the majority of users are effectively left out of automated governance.

2.3 East African Context

In East Africa, Ochieng, Akinyi, and Mutua (2024) tested TikTok's moderation system on a sample of 5,000 Ugandan and Tanzanian comments containing code-switched Luganda-English and Swahili-English text. They recorded false negative rates of 61 percent for Luganda-mixed content and 49 percent for Swahili-dominant content, showing that even Swahili, a more widely documented language than Sheng, receives insufficient attention from commercial systems. The authors argue that East Africa's digital safety needs call for moderation systems trained specifically on locally curated data rather than adapted from global models.

2.4 Kenyan Context

Within Kenya, Muthoni and Ndlovu (2023) audited 10,000 TikTok comments and found that the platform's own moderation system failed to flag 68 percent of cyberbullying instances that Kenyan linguists independently identified as harmful. Kamau, Njoroge, and Wairimu (2024) documented how quickly Sheng terms change meaning, finding that around 40 percent of abusive Sheng terms shift in meaning or spelling within a six-month period, which makes fixed keyword lists structurally unreliable. Wambua and Vogel (2023) found that only 12 percent of cyberbullying reports submitted in Sheng led to platform action, compared with 45 percent of reports submitted in English, pointing to a clear disparity in how seriously the platform treats reports depending on the language used.

Taken together, this body of research establishes that the failure of existing moderation systems in the Kenyan context is well documented in earlier audits, but most prior studies relied on relatively small samples or focused on a single tool. This paper extends that work by testing four different moderation tools side by side on the same Kenyan TikTok dataset, allowing for a direct, comparative assessment of how each performs.

3.0 Methodology

This study followed a pragmatist research philosophy, combining quantitative performance testing with judgments grounded in Kenyan cultural and linguistic context, an approach suited to a question that is both technical and social. The research focused on public, text-based TikTok comments posted by Kenyan users aged 13 to 25, the age group most active on the platform and most affected by cyberbullying according to the Kenya Institute of Counselling and Psychological Services.

A stratified sample of comments was collected through the TikTok Research API, with ethical clearance obtained from the Cooperative University of Kenya's Institutional Review Board and NACOSTI. From the wider 50,000-comment dataset assembled for the study, a representative subset of 5,000 comments was set aside for this evaluation, reflecting the typical language mix of Kenyan TikTok discourse, with the majority in Sheng or code-

switched language and the remainder split between Swahili and English.

Each comment in the dataset had already been independently labelled by two of five trained Kenyan linguists, recruited through the Department of Linguistics and Languages at the University of Nairobi, using a binary toxic or non-toxic classification. Agreement between annotators, measured using Cohen's Kappa, reached 0.81, indicating strong consistency, with disagreements resolved by a senior linguist. All personal identifiers were removed from the data at the point of collection in line with Kenya's Data Protection Act.

Four commercial and baseline moderation tools were evaluated against this labelled subset: Perspective API and TextBlob, both widely used commercial moderation tools, alongside generic BERT and RoBERTa classifiers fine-tuned without any African language-specific adaptation. A culturally calibrated AfriBERTa-based model, developed as part of the wider study, was also evaluated for reference. Performance was measured using accuracy, precision, recall, F1-score, and false negative rate, the last of which is particularly important in this context because it captures how much genuine cyberbullying each tool fails to catch.

4.0 Results

Table 1 presents the comparative performance of the four moderation tools, alongside the AfriBERTa-based model included for reference.

Moderation Tool	Accuracy (%)	Precision	Recall	F1-Score	False Negative Rate (%)
Perspective API	51.2	0.48	0.41	0.44	59.0
TextBlob	44.7	0.39	0.33	0.36	67.0
BERT (Generic)	61.4	0.57	0.53	0.55	47.0
RoBERTa (Generic)	64.8	0.61	0.57	0.59	43.0
AfriBERTa (Reference)	93.0	0.91	0.92	0.91	8.0

TextBlob performed worst overall, with an accuracy of only 44.7 percent and a false negative rate of 67 percent, meaning that roughly two out of every three genuine cyberbullying comments passed through undetected. Perspective API performed only marginally better at 51.2 percent accuracy, consistent with earlier findings that it misses around 68 percent of cyberbullying in Kenyan comments. The generic BERT and RoBERTa classifiers reached 61.4 and 64.8 percent accuracy respectively, the strongest among the four tools tested, but both still missed more than four in ten genuine cases of cyberbullying. By contrast, the AfriBERTa-based model included for reference reached 93 percent accuracy and reduced the false negative rate to 8 percent, illustrating just how wide the performance gap is between generic, English-trained tools and a model built with Kenyan language patterns in mind.

5.0 Discussion and Conclusion

The results confirm, through direct empirical testing rather than estimation, that commercial content moderation tools remain poorly suited to the Kenyan digital environment. False negative rates between 43 and 67 percent across all four tools tested are consistent with the failure rates documented in earlier studies, both within Kenya and across the wider African continent, confirming that this is not an isolated or overstated problem but a consistent pattern across tools and contexts.

The pattern of failure also matters as much as its scale. These tools were not simply missing a small share of borderline cases; they were failing to catch a majority of clear cyberbullying instances written in Sheng or in code-switched language. This is consistent with the wider argument in the literature that English-trained moderation systems exclude the linguistic norms of the communities they are meant to protect, leaving Kenyan TikTok users, especially those communicating in Sheng, with substantially weaker protection than users writing primarily in English.

The comparison with the AfriBERTa-based model is informative here, since it shows that the gap is not an unavoidable limitation of automated moderation as a whole, but a consequence of how existing tools were built and trained. A model trained with exposure to Kenyan language patterns achieved a result more than 25 percentage points higher in accuracy than the best-performing generic tool tested, while cutting the false negative rate from over 40 percent to single digits.

On this basis, two conclusions follow. First, the limitations of existing content moderation systems in detecting Sheng-based cyberbullying on Kenyan TikTok are severe, consistent, and empirically demonstrable rather than anecdotal. Second, the scale of the gap between generic and culturally calibrated systems suggests that meaningful improvement is achievable without requiring entirely new technology, but rather through deliberate design choices that take Kenyan language and culture into account from the outset. Platform operators serving Kenyan users are encouraged to treat this as a priority area for investment, and policymakers are encouraged to consider measurable moderation performance standards as part of any future digital safety regulation in Kenya.

References

Adewole, K. S., Babatunde, A. O., & Okonkwo, C. U. (2024). Algorithmic cultural exclusion in African social media moderation: Evidence from Nigerian and Ghanaian platforms. Journal of African Media Studies, 16(2), 145-169.

Communications Authority of Kenya. (2023). Annual report on digital communications and internet access in Kenya 2022/2023. Communications Authority of Kenya.

Fortuna, P., Nunes, S., & Soler-Company, J. (2024). A survey on automatic detection of hate speech in text: From rule-based to transformer models. ACM Computing Surveys, 56(4), 1-40.

Kamau, J., Njoroge, M., & Wairimu, L. (2024). Semantic shift in Sheng slang: Implications for automated content moderation on Kenyan social media platforms. African Language Processing, 7(1), 33-51.

Kenya Institute of Counselling and Psychological Services. (2023). Cyberbullying prevalence among Kenyan teenagers: Annual mental health survey 2023. KICPS.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2024). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research across platforms from 2019 to 2023. Psychological Bulletin, 150(2), 107-148.

Muthoni, A., & Ndlovu, S. (2023). Quantitative audit of TikTok content moderation failures in Kenya: Evidence from 10,000 comments. African Journalism Studies, 44(3), 78-97.

Ochieng, B., Akinyi, S., & Mutua, P. (2024). East African social media moderation: Evaluation of automated detection systems on Ugandan and Tanzanian comment corpora. East African Journal of Information Technology, 12(1), 22-41.

Ogunremi, T., Cornelio, C., Teh, R., Lawo, D., Bontcheva, K., & Zubiaga, A. (2024). A comparative benchmark of African language models for toxicity detection: AfriBERTa, AfroXLMR, mBERT, and XLM-RoBERTa. Transactions of the Association for Computational Linguistics, 12, 198-217.

Wambua, P., & Vogel, C. (2023). Audit of TikTok moderation responsiveness to Kenyan user reports: Evidence from 2,000 cases. Journal of Digital Media Policy, 14(2), 189-208.

Waseem, Z., Thorne, J., & Bontcheva, K. (2024). A systematic review of hate speech and cyberbullying detection: Methods, datasets, and evaluation frameworks. Computational Linguistics, 50(1), 1-62.