BRIEFING July 2022

FACEBOOK UNABLE TO DETECT HATE SPEECH WEEKS AWAY FROM TIGHT KENYAN ELECTION

Despite the risk of violence around the upcoming Kenyan election, Global Witness' new investigation conducted in partnership with legal non-profit Foxglove finds Facebook appallingly failed to detect hate speech ads in the two official languages of the country: Swahili and English. This follows a similar pattern we uncovered in Myanmar and Ethiopia, but for the first time also raises serious questions about Facebook's content moderation capabilities in English. Facebook itself has praised its "super-efficient AI models to detect hate speech"[1] but our findings are a stark reminder of the risk of hate and incitement to violence on their platform. Even worse, in the lead up to a high stakes election, this is a time you would expect Facebook's systems to be even more primed for safety.

"In the backdrop of elections, it is even more important for us to detect potential hate speech and prevent it from spreading."[2]
Meta, 2022

On the 9th of August Kenyans will go to the polls for general elections, which are expected to be tightly contested and bitterly fought. While the situation in Kenya has improved in many respects, it remains a volatile political landscape and the risks are real. Given Kenya's recent history of electoral violence and the "polarised, ethnically driven and personalist politics" of the country, it remains vulnerable to unrest.[3] Some of the worst violence occurred after the 2007 elections, when tribal tensions were laid bare after inflammatory electoral campaigns and a disputed result. As many as 1,300 people were killed and hundreds of thousands fled their homes.

Voters queuing at a polling station in Kiambu, Kenya, in 2017, as polls opened for presidential elections. Photo credit: © SIMON MAINA/AFP via Getty Images

OUR INVESTIGATION

We decided to test Facebook's ability to detect hate speech ahead of the Kenyan elections, sourcing ten real-life examples of hate speech used in Kenya since 2007 and submitting them for approval. In total we submitted twenty ads[4] to Facebook, which covered the ten real-life hate speech examples and their corresponding translation in English or Swahili. By conducting the investigation in Kenya Global Witness was able to legitimately test for the first time whether Facebook might be better at detecting English language hate speech over Swahili, given these are the two official languages of the country.

> "We're a pioneer in artificial intelligence technology to remove hateful content at scale."[5]
> Nick Clegg, VP of Global Affairs and Communications at Facebook, 2020

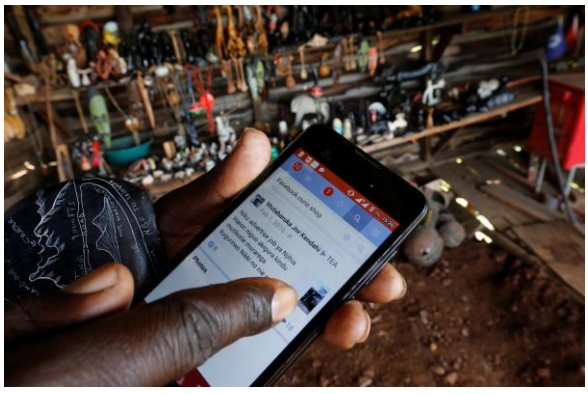
Much to our surprise and concern, all hate speech examples in both languages were approved, with one exception: our English language hate speech ads were initially rejected for failing to comply with Facebook's Grammar and Profanity policy. Facebook invited us to update the ads, and after making minor corrections they were similarly accepted. Seemingly our English ads had woken up their AI systems, but not for the reason we expected.

All of the ads we submitted violate Facebook's Community Standards[6], qualifying as hate speech and ethnic-based calls to violence. Much of the speech was dehumanising, comparing specific tribal groups to animals and calling for rape, slaughter and beheading. We are deliberately not repeating the phrases used here as they are highly offensive.[7]

We submitted the hate speech in the form of ads, as this enables us to remove them before they go live, while still being reviewed by Facebook and undergoing Facebook's content moderation process. According to Facebook, this often includes proactive review using automated and manual tools.[8] Facebook lauds its own system for reviewing ads with advertisers being held to an 'even stricter' standard.[9]

When asked for comment on the investigation findings, a Meta spokesperson - Facebook's parent company - responded to Global Witness[10] that they've taken "extensive steps" to help Meta "catch hate speech and inflammatory content in Kenya" and that they're "intensifying these efforts ahead of the election". They state Meta has "dedicated teams of Swahili speakers and proactive detection technology to help us remove harmful content quickly and at scale". Meta acknowledges that there will be instances where they miss things and take down content in error, "as both machines and people make mistakes."

After we alerted them to our investigation, Meta then put out a new public statement[11] on its preparations ahead of the Kenya elections - specifically highlighting their apparent action taken to remove hateful content in the country - we then submitted two more ads to see if there had indeed been any improvement in Facebook's detection of hate speech ads. Once again the ads we resubmitted in Swahili and English were approved.



People in Kenya are spending more time on Facebook. Photo credit: © REUTERS/Thomas Mukoya

ONLINE HATE SPEECH IN KENYA

Since 2007 more and more people in Kenya are spending time on social media, with Facebook having over 12 million users - over 20% of the Kenyan population.[12] Reports of online hate speech and disinformation have been growing, including in relation to the 2017 elections on Facebook, WhatsApp and Twitter and most recently on TikTok.[13]

Earlier this year a Kenyan Government agency tasked with addressing inter-ethnic violence found that hate speech on social media platforms in 2022 had increased 20 percent and warned of their misuse to spread "ethnic hate speech and incitement to violence".[14] Facebook was found to spread the most hate and incitement, followed by Twitter.[15]

WHAT NEEDS TO CHANGE

All of this points to a broken system. For one of the world's wealthiest companies, with staggering reach and a responsibility not to facilitate division and harm, Facebook can and should do better.

In 2020 following advertiser pressure during the #StopHateForProfit boycott, CEO Mark Zuckerberg said Facebook was going to do much more to tackle hate on its platform - including widening what is considered 'hateful' in ads.[16] But our repeated findings - in Myanmar, Ethiopia and now Kenya - raises serious questions about whether these commitments were followed through, particularly in all parts of the world. This also follows reports from employees that Zuckerberg is no longer prioritising safeguarding elections, instead focusing on the so-called 'metaverse' - Meta's new frontier of growth.[17]

Importantly, this is not the fault of the individual content moderators, who all too often are asked to undertake deeply traumatising work - including in Kenya - with scant regard for their mental health and decent working conditions. Earlier this year a former moderator at Facebook filed a lawsuit in Kenya against Meta and its local outsourcing company Sama, alleging and seeking reforms to poor working conditions - including "irregular pay, inadequate mental health support, union-busting, and violations of their privacy and dignity".[18]

While the EU is taking a lead globally to regulate Big Tech companies and force meaningful oversight - including requiring the platforms to assess and mitigate the risk that their services allow hate speech to flourish - platforms should also be acting of their volition to protect their users fully and equally.

> "At Meta, we know we have an important responsibility when it comes to helping people participate in elections and to ensure safe, secure, and free elections"[19]
>
> Meta, 2022

We call on Facebook to:

- Urgently increase the content moderation capabilities and integrity systems deployed to mitigate risk before, during and after the upcoming Kenyan election.
- Properly resource content moderation in all the countries in which they operate around the world, including providing paying content moderators a fair wage, allowing them to unionise and providing psychological support.
- Routinely assess, mitigate and publish the risks that their services impact on people's human rights and other societal level harms in all countries in which they operate.
- Publish information on what steps they've taken in each country and for each language to keep users safe from online hate.

ENDNOTES

[1] https://ai.facebook.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/

[2] https://about.fb.com/news/2022/02/how-meta-is-prepared-to-protect-the-upcoming-state-elections-in-india/

[3] https://www.crisisgroup.org/africa/horn-africa/kenya/b182-kenyas-2022-election-high-stakes

[4] The ads were in the form of an image (text on a plain background) and were not labelled as being political in nature.

[5] https://about.fb.com/news/2020/07/facebook-does-not-benefit-from-hate/#:~:text=We're%20a%20pioneer%20in,faster%20than%20YouTube%20and%20Twitter.

[6] https://transparency.fb.com/en-gb/policies/community-standards/

[7] Researchers interested in knowing the exact wording of the hate speech examples we used are welcome to request this from us by writing to digitalthreats@globalwitness.org

[8] https://www.facebook.com/business/help/2001034850142726

[9] https://www.facebook.com/business/about/ad-principles

[10] "We've taken extensive steps to help us catch hate speech and inflammatory content in Kenya, and we're intensifying these efforts ahead of the election. We have dedicated teams of Swahili speakers and proactive detection technology to help us remove harmful content quickly and at scale. We've also created a team of subject matter experts working on the election, including individuals with expertise in misinformation, hate speech, elections and disinformation. Despite these efforts, we know that there will be examples of things we miss or we take down in error, as both machines and people make mistakes. That's why we have teams closely monitoring the situation and addressing these errors as quickly as possible." – a Meta spokesperson

[11] https://about.fb.com/news/2022/07/how-metas-preparing-for-kenyas-2022-general-election/

[12] https://www.statista.com/statistics/1029198/facebook-user-share-in-kenya-by-age/

[13] https://qz.com/africa/1033181/whatsapp-and-facebook-are-driving-kenyas-fake-news-cycle-ahead-of-august-elections/ ; https://foundation.mozilla.org/en/campaigns/kenya-tiktok/

[14] https://cohesion.or.ke/images/docs/downloads/MEDIABRIEF.pdf

[15] https://www.capitalfm.co.ke/news/2022/04/ncic-says-facebook-leaders-in-hate-speech-inflammatory-remarks/

[16] https://www.facebook.com/zuck/posts/10112048980882521

[17] https://www.nytimes.com/2022/06/23/technology/mark-zuckerberg-meta-midterm-elections.html

[18] https://www.theguardian.com/technology/2022/may/10/ex-facebook-moderator-in-kenya-sues-over-working-conditions

[19] https://www.facebook.com/gpa/blog/kenya-preparing-for-the-2022-elections