BUILD UP

THE KENYA POLARIZATION FOOTPRINT

a measure of societal divides on social media in Kenya

The Kenya Polarization Footprint measures how social media is dividing Kenyans. Specifically, we measure "affective polarization", where people dislike and distance themselves from others based on their identity. This is not the same as a difference of opinion: differences of opinion are good for democratic debate, and for our diverse society. Identity-based polarization is not: increasing dislike, distrust, and animosity towards other social groups results in societal divides that can eventually lead to violence.

The way content is shared on social media makes societal divides wider. Measuring just how much online polarization is present on each social media platform is an important first step to defending our society.

"The Kenya Polarization Footprint reveals deepening societal divides driven by affective polarization on social media where identity-based hostility fuels distrust and animosity. This growing digital harm is largely enabled by the presence of super users who perpetuate a cycle of division and platforms receiving perverse incentives from polarised content further fragmenting public discourse and posing a tangible threat to Kenya's social fabric. The report underscores that polarization is not merely a digital nuisance but a measurable social harm. Addressing it requires a shift in perspective from individual responsibility to collective action. It demands platform accountability, participatory governance, and systemic interventions to safeguard democratic discourse and social cohesion." – Rachel Olpengs, Lead Coordinator, National Coalition on Freedom of Expression and Content Moderation in Kenya (FECoMo)

METHOD OVERVIEW

We contacted 5000 Kenyans, chosen to represent the make-up of the Kenyan population, and measured their online experience in two ways:

1. By **observing** their social media feeds on TikTok, YouTube, Instagram, Facebook and X (Twitter), and applying three measurements of polarization to this content.
2. By **asking** them about their experience with positive and negative content on social media over the past 4 weeks, and how it impacted them.

For details of how exactly we conducted this measurement, we have made our full methodology publicly available here.

KENYA POLARIZATION FOOTPRINT RANKING

Overall, **we rank the affective polarization of the five platforms** we looked at as follows:

Platform	Facebook	Instagram	TikTok	YouTube	X (Twitter)
Polarization footprint ranking: 1 is best (lowest polarization) and 5 is worst (highest polarization)	1	2	3	4	5

How did we calculate this? The polarization footprint is a composite measure made up of three parts, each measuring a component of affective polarization – attitude polarization, norm polarization and interaction polarization.

- **Attitude polarization** looks at how much language used in social media posts and comments expresses negativity towards an identity group, which we call "othering" (more othering = more polarization);
- **Norm polarization** looks at how often people challenge polarized attitudes in the comments in a social media thread (no challenge = polarization is "normal");
- **Interaction polarization** looks at how much users of a social media platform are fragmented into dissimilar clusters (more fragmentation = more polarization).

We rank platforms for each type of polarization, and then combine these rankings into a single overall ranking. For example, if platform A was observed to have the 2nd lowest levels of attitude polarization, the 3rd lowest levels of norm polarization, and the lowest degree of interaction polarization, its overall score would be 6 (2 + 3 + 1). By ranking the platforms by this score, we produce an overall ranking of the platforms by their level of affective polarisation.

What does the polarization footprint mean? Across all measures of polarization that impact societal divides, X (Twitter) does worst. This means that when you scroll through content on X, you are most likely to become polarized.

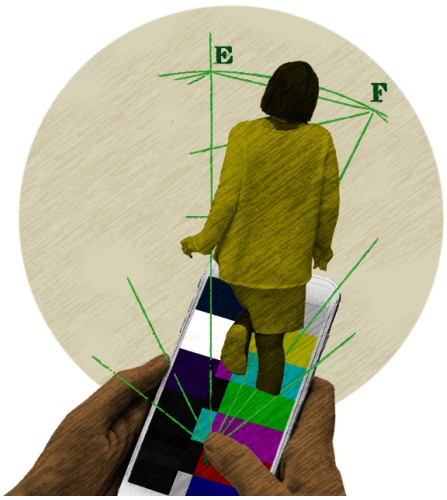
The other platforms do better, but still have a significant impact on divides. The scores and ranking for each type of polarization help us unpack this a bit further. For example, Facebook and TikTok have a middling ranking for attitude and norm polarization, but on Facebook the interaction polarization score is much lower, meaning that people's content networks are less fragmented. That ends up making Facebook the least toxic platform, with TikTok ranking third most toxic.

Platform	Attitude		Norm		Interaction		Footprint	
	Score	*Rank*	*Score*	*Rank*	*Score*	*Rank*	*Score*	*Rank*
Facebook	0.52	2	74.58	3	26.9	1	6	1
Instagram	0.36	1	77.66	4	31.64	2	7	2
TikTok	0.53	3	73.6	2	35.04	3	8	3

YouTube	0.62	4	66.04	1	48.94	4	9	4
X (Twitter)	3.83	5	83.02	5	62.48	5	15	5

Comparing YouTube and Instagram also offers some interesting insights: Instagram ranks best out of all the platforms on attitude polarization, but is the second to worst on norm polarization, meaning there are few challenges to polarized attitudes when they are expressed in an Instagram post or comment. YouTube is the exact opposite: it ranks second to worst on attitude polarization, but has the least norm polarization – meaning attitude polarization is more often challenged. It's again the interaction polarization score that impacts the final ranking: Instagram's network is not very fragmented; YouTube's is, making it the second to worst overall.

To get into the details of what is happening in each platform, the rest of this report disaggregates these calculations, and offers additional insights.



Attitude Polarization Scores

Attitude polarization looks at whether the language used in social media posts and comments expresses any of the following about an identity group: negative stereotypes, dehumanization, deindividuation, vilification, or calls to violence. The attitude polarization score is the percentage of posts and comments that contain this language.

The table below details results by platform, as % of recommended threads and as time in a day, where a "thread" is a post and all the comments beneath it. Our measures are deliberately designed to show the **minimum % of recommended threads that are polarizing**. It's a minimum measure both because we were very conservative in our annotation (8 people annotated all polarizing content, and we used a "tie-down" average), and because our measure assumes that no other comment in a post-comments thread is polarizing (when in

fact we know that it's likely that once there is one polarizing post or comment, there are likely more in the same thread, which would make the overall % of content that is polarizing higher).

To give one example, we expect a minimum of 3.81% of threads on X (Twitter) is polarizing. This means that for every hour you spend on X, on average you're spending a minimum of just over 2 minutes on polarizing content. 55% of Kenyans are spending 3 - 9 hours on social media per day; if that time were spent on X, then your average Kenyan would spend at least 7 to 21 minutes seeing polarization[1]. Imagine breathing from a car exhaust pipe for 7 to 21 minutes every day. Polarizing content is making us sick. Imagine being shouted at for 7 to 21 minutes every day. No wonder we feel angry.

"Imagine breathing from a car exhaust pipe for 7 to 21 minutes every day. Polarizing content is making us sick."

What's worse, some Kenyans (5%) spend more than 12 hours on social media per day, which means they are breathing in polarization for a minimum of 27 minutes every day; getting shouted at for a minimum of 27 minutes every day. These super users are often also the most prolific on social media – the influencers, the rabble rousers, the ones who drive narrative. The more they see, the angrier they get, the more toxicity that goes back out into our society. It's a vicious cycle.

Platform	**X (Twitter)**	**Facebook**	**YouTube**	**TikTok**	**Instagram**
minimum % of recommended threads with attitude polarization (in brackets, the 95% confidence interval)	3.83% (3.07, 4.78)	0.52% (0.28, 0.95)	0.62% (0.35, 1.08)	0.53% (0.29, 0.97)	0.36% (0.17, 0.73)
What does that mean in terms of time spent[2] by Kenyans seeing content that increases societal divides? *(Note: careful, this table assumes someone spends their time only on one platform, but we know they probably jump around!)*					
31% of Kenyans spend 1 - 3 hours on social media per day, so minutes seeing polarization...	2 to 7 minutes per day	0.3 to 1 minutes per day	0.4 to 1 minutes per day	0.3 to 1 minutes per day	0.2 to 0.7 minutes per day

[1] Our time measures are likely also conservative, because they assume we spend the same amount of time on every piece of content, when it seems likely we would spend more on things that are more emotionally charged, such as polarizing content.

[2] Time spent taken from: https://www.geopoll.com/blog/smartphone-and-social-media-usage-2025/#:~:text=2023%20data%20collection-,Time%20spent,invest%20in%20these%20online%20interactions

55% of Kenyans spend 3 - 9 hours per day on social media per day, so minutes seeing polarization...	7 to 21 minutes per day	1 to 3 minutes per day	1 to 3 minutes per day	1 to 3 minutes per day	0.7 to 2 minutes per day
9% of Kenyans spend 9 - 12 hours on social media per day, so minutes seeing polarization...	21 - 27 minutes per day	3 to 4 minutes per day	3 to 5 minutes per day	3 to 4 minutes per day	2 to 3 minutes per day
5% of Kenyans spend more than 12 hours on social media per day, so minutes seeing polarization...	>27 minutes per day	>4 minutes per day	>5 minutes per day	>4 minutes per day	>3 minutes per day

Norm Polarization Scores

So is anyone challenging this polarization? Norm polarization is at play when polarized attitudes go unchallenged and people come to expect they are normal on social media. These combative norms of interaction reinforce the erosion of trust between people, and make affective polarization worse. The norm polarization score is the percentage of threads where there is no challenge to polarized attitudes in the comments in a social media thread where there is attitude polarization. As with the attitude polarization scores, our measure is conservative and reports the **minimum % of polarizing threads where there is no challenge**.

"Our results show that the most common experience is that no-one challenges attitude polarization on social media."

Platform	**X (Twitter)**	**Facebook**	**YouTube**	**TikTok**	**Instagram**
minimum % of polarizing threads where no-one challenges the polarization	83.02%	74.58%	66.04%	73.6%	77.66%

Our results show that across all platforms, the most common experience is that no-one challenges attitude polarization. For example, on Instagram, when someone posted something that was polarizing, 78% of the time no-one challenged it. Even on YouTube, the platform with the lowest norm polarization score, 66% of the time no-one challenges polarization in a thread. No challenge signals a norm shift towards negative expectations — that is, it signals that people think attacks on others are normal. Imagine that you are being shouted at, and no by-stander steps in to defend you or to lower the tone. That's what norm polarization is about.

Interaction Polarization Scores

There is another aspect of affective polarization that makes things even worse: we are all experiencing different versions of it. Interaction polarization is the extent to which users of a social media platform are fragmented into dissimilar clusters, which impacts both the interests and affiliations of people, creating a self-reinforcing cycle of polarization.

Whereas we measure attitude and norm polarization based on individual behaviors (posting content, reacting to content), interaction polarization is a network-wide dynamic, not a post-level attribute. We measure interaction polarization as the degree of fragmentation in the accounts people see content from in their feed – that is, how easy it is to predict based on one account seen what other accounts a user is likely to see in their feed.

"When interaction polarization is high, narratives, experiences and truth in our society become fragmented."

Platform	X (Twitter)	Facebook	YouTube	TikTok	Instagram
Measure of fragmentation in what accounts people see in their feeds (higher score = more polarization)[3]	62.48	26.9	48.94	35.04	31.64

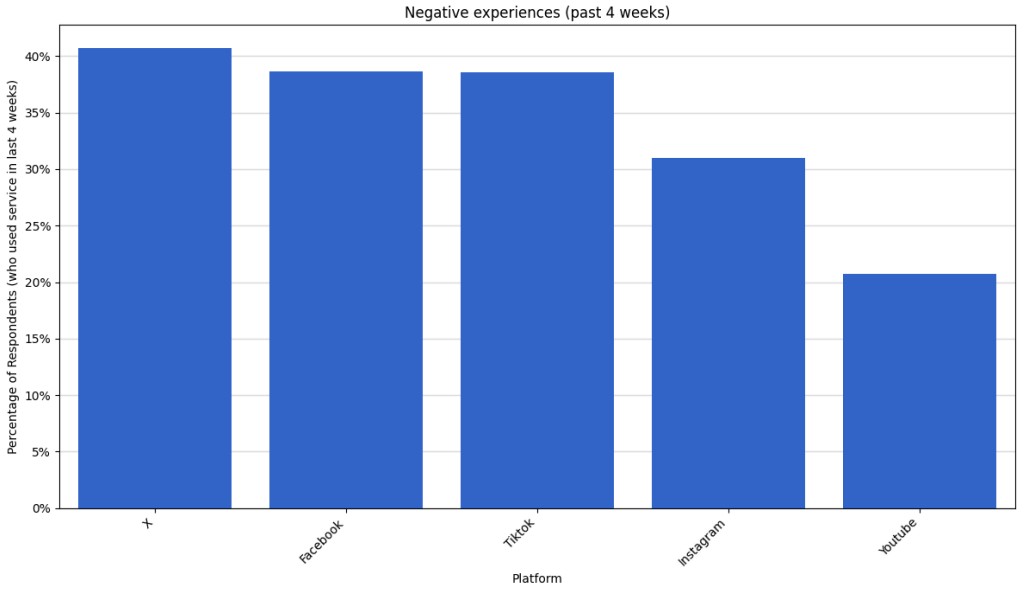
The higher the measure, the more fragmentation of narratives, experiences and truth in our society – and these are important to social cohesion. The scores above tell us that – especially on X and YouTube – people live in their own social media content universes separate from the universes of others.

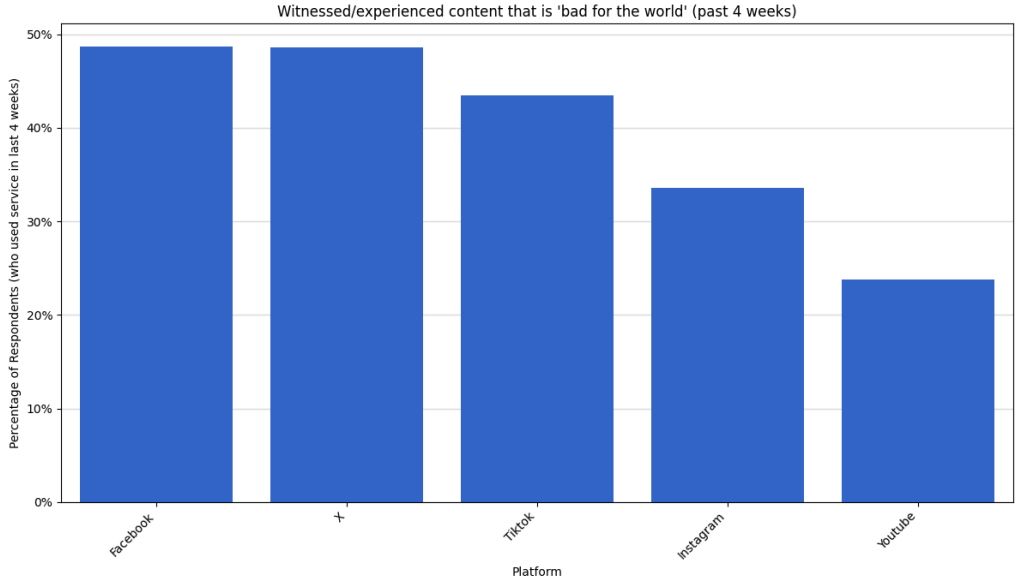
[3] Specifically, total correlation in the joint distribution with Bernoulli marginals indicating whether or not a participant was recommended content from each of the top 100 most-recommended accounts.

KENYAN'S EXPERIENCE OF POLARIZATION

Observing what polarizing language, challenges to this language, and patterns of interaction look like on social media platforms is an important measure of how these platforms are affecting our society. But we worried that we wouldn't get the full picture of the impact this is having on Kenyans unless we asked directly about their experiences on social media.

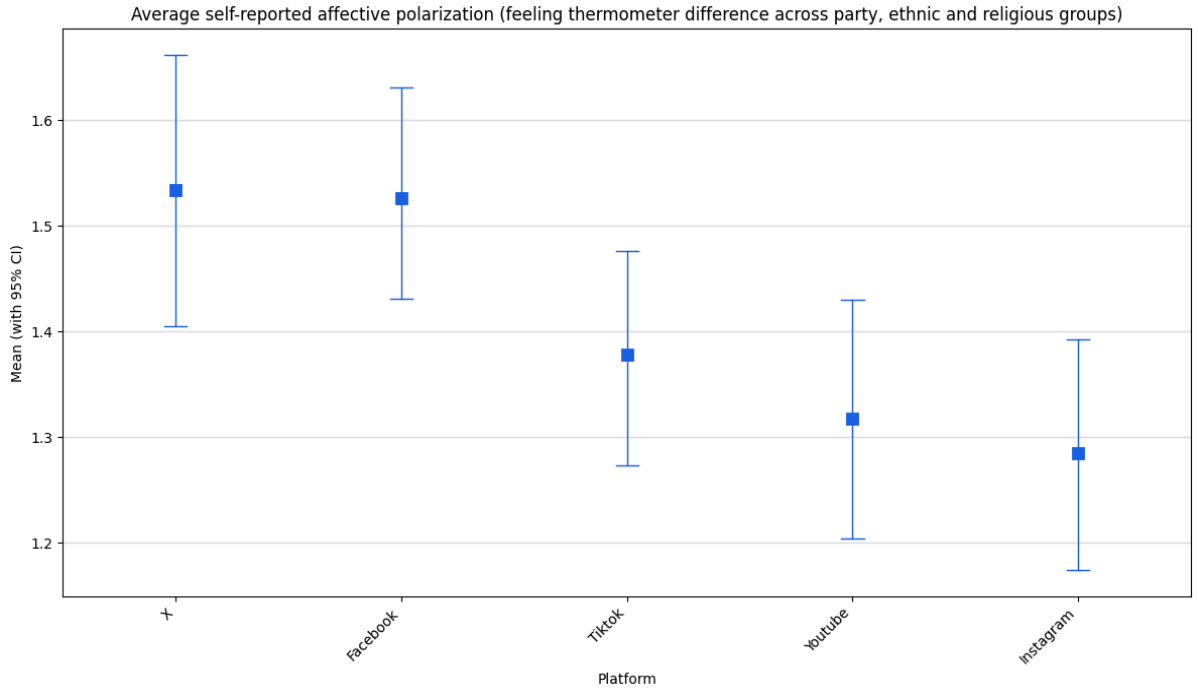
We asked 5000 Kenyans, chosen to represent the make-up of the Kenyan population, whether, in the past 4 weeks, they had negative experiences on a social media platform or saw content they thought was "bad for the world". On all platforms except YouTube, over 30% of the people we surveyed had a direct negative experience. The same was true of content they considered to be "bad for the world". The results for X (Twitter) track with what we would expect from the polarization footprint; the results for YouTube and Facebook do not track with their polarization footprint.





"On all platforms except YouTube, over 30% of the people we surveyed had a direct negative experience."

What's more, we also asked a series of questions about how they felt towards people of their same political party, ethnicity and religious group, relative to how they felt towards people from other parties, ethnicities or religious groups. We call this their self-reported affective polarization. We found that people using X (Twitter) and Facebook both had higher self-reported polarization than those using the other platforms.



These results tell us one thing clearly: X (Twitter) is a toxic platform, across all our measures. They also leave us with some questions: what is happening on Facebook and YouTube for people to self-report so differently to what we observe in their feeds? We'll be taking that question forward with us, and conducting some further analysis of the survey results to try to unpack what is going on.

WHAT NEXT?

We set out to measure affective polarization on social media in Kenya because we are worried that disagreements online spiral out of control, and that this happens because they become about people rather than ideas. Increases in affective polarization are a digital harm that drives conflict and harms democracy.

Research on digital platforms shows that content that has affective polarization does something else: it captures attention online, it increases engagement. That's why social media companies that rely on capturing attention as a way to make money have no incentive to control for polarization.

"What happens if we understand polarization as a negative externality – a pollution we can measure?"

What happens if we understand polarization not only as a digital harm that drives conflict, but also as a negative externality – a pollution we can measure? That's what the polarization footprint does. Now that we have that measure, can we find ways to clean it up? We're starting a series of conversations about this, and we invite you to join us. Follow us on LinkedIn or contact team@howtobuildup.org for more information.