Community Notes as One Layer in a Multi-Stack Moderation System and Why Fact-Checking Remains Critical for Information Integrity In India

Submitted by The Trusted Information Alliance, India

Argument Summary

Community Notes (CN) is an addition to Meta's integrity toolbox, offering community participation, and contextual explanations around disputed content. However, CN's crowdsourced design also carries structural limitations that make it insufficient as a standalone solution in India's diverse and high-stakes information environment. This is also true for many other countries and regions in the world, and especially crucial for India as it is the largest market for Meta's products.

As smartphone penetration deepens and data costs remain low, a new wave of users from tier-2, tier-3 cities and rural areas are accessing content in Hindi, Tamil, Telugu, Bengali, Marathi, and other local languages. With 22 officially recognized languages under the Indian Constitution's Eighth Schedule and 121 languages with significant user bases, India is home to an extraordinary linguistic diversity. While this linguistic richness has historically coexisted with a public sphere often dominated by English and Hindi, particularly in traditional media and early internet phases, with the next wave of internet users emerging from non-metro regions, language is no longer a barrier to participation. Instead, it is the medium of choice.

In addition to this, take into account India's low digital literacy with only around 38% of the households reported to be digitally literate. This makes the majority of the non-English speaking new internet users more vulnerable to misinformation.

Our analysis of Community Notes on X's platform, shows that since its launch, only about 2,200 notes have been published in Indian languages, written by only 805 unique contributors, as of December 2025. Of these, just 55 notes in Indian languages were rated as "currently helpful", as of December 6, 2025. (Refer data in Table 1 and 2 below)

Independent studies examining Meta's Community Notes reflect similar patterns of severe linguistic imbalance. Because of high consensus thresholds, notes in Indian languages rarely get published, leaving large regional-language spaces underserved.

Moreover, studies tell us that CN's operational design is well-suited for correcting simple factual inaccuracies but is not equipped to evaluate complex or harm-based narratives such as gendered, communal, or caste-driven misinformation where trained experts are essential.

Moreover, CN's reliance on cross-group agreement significantly delays visibility (averaging 14 days globally), while notes that reference professional fact-checkers appear earlier, are more trusted, and achieve higher visibility. For these reasons, Community Notes should function as one layer within a multi-stack approach, complemented, not replacing third-party fact-checking. This analysis draws on independent academic and institutional studies of X's Community Notes model, which informs Meta's CN architecture, given the limited India-specific transparency data available from Meta platforms.

Key Arguments Against Replacing Third-Party Fact-Checking With Community Notes Model In India

South Asian languages are severely underrepresented, creating an information integrity gap. South Asian languages (Hindi, Urdu, Bengali, Tamil, etc.) account for only **0.094% of the global CN corpus**, leaving vast vernacular online spaces vulnerable to unchecked misinformation. The CN interface, guidelines, and sign-up process are still heavily English-centric. (CSOH, 2025).

The TIA also extracted data on X's Community Notes to validate the share of Indian languages. As per our data, more than 25 lakh notes were written since the inception of Community Notes till December 5, 2025 and only 0.894% were written in Indian languages. Only 55 notes were rated as 'helpful'.

Table 1: Share Of Notes In Indian languages In Community Notes Published on X In India, since inception

Labels	Currently rated helpful	Currently not rated helpful	Needs more ratings	Blank	Grand Total	Share in total CN
Bengali			9	1	10	0.0004%
Gujarati	1		7	2	10	0.0004%
Hindi	40	11	1063	150	1264	0.0504%
Kannada			8	5	13	0.0005%
Malayalam			58	25	83	0.0033%
Marathi			58	12	70	0.0028%
Punjabi				2	2	0.0001%
Tamil	4	2	193	25	224	0.0089%
Telugu	1	1	18	1	21	0.0008%
Urdu	9	3	514	18	544	0.0217%
Grand Total	55	17	1928	241	2241	0.0894%

The consensus algorithm fails in low-volume language communities. Notes in South Asian languages face **structural barriers** leading to a publication rate of only 2.30% compared to 8.25% globally. This percentage is lower for Indian languages, as shown in Table 1 above. High-quality notes often remain **"stuck in limbo"** because low reviewer density prevents the bridging algorithm from finding the necessary **cross-cluster agreement**. This requires vernacular notes to clear a **substantially higher threshold of review** (e.g., typically needing over 80 ratings to publish). (CSOH, 2025)

Increased participation does not guarantee success, leading to bottlenecks. The system fails to scale efficiently; as seen globally, even where contributor volume increases, publication success can be low (Spanish notes still lag) or cause new bottlenecks (too many English notes for raters to choose from), resulting in fewer notes breaking through. This undermines the promise of collective moderation. (DIA 2025)

Table 2: Number of Unique Contributors In CNs In Indian Languages Since Inception Till Dec 6, 2025

Language	Unique Contributors	Share In Total
Bengali	8	0.002%
Gujarati	3	0.001%
Hindi	389	0.107%
Kannada	10	0.003%
Malayalam	26	0.007%

Marathi	45	0.012%
Punjabi	2	0.001%
Tamil	100	0.028%
Telugu	18	0.005%
Urdu	204	0.056%
Grand Total	805	0.222%

CN is far too slow to counter viral harms, particularly during crises. Despite improvements, the average time for a note to go public is still **14 days**. This is insufficient to counter online harms that spread within **hours** (DIA 2025). Furthermore, CN usage in India surges only *reactively* during high-stakes events like the General Election, meaning **hundreds of drafts languish during the weeks voters need them.** (CSOH, 2025)

The system is vulnerable to partisan misuse in polarized environments. In South Asia, a significant portion of notes carry **political biases** (46.0%) or **potentially harmful/divisive tone** (4.5%). Examples include notes using personal slurs ("Italian-mafia supporters") or political rhetoric rather than fact-checking. This creates the risk of **"crowdsourced propaganda,"** leading users to distrust the feature as merely another arena for political mud-slinging (CSOH, 2025).

Relevance of Fact-Checks

Expert fact-checking provides essential speed and quality indicators. Community notes that **cite evidence from fact-checking organizations** (the third most used global reference) are significantly **more likely to become visible** (12% visibility rate vs. 8.3% overall). Crucially, Community Notes that cite a fact-checker are considered more useful by users and appear on tweets more quickly—90 minutes earlier than general notes (Maldita 2025).

Expertise is crucial for nuance and local context. Experts, such as journalists and fact-checkers, are **"deeply attuned to local languages, cultures and socio-political nuances,"** and are important for assessing the **"check-worthiness"** of narratives and their potential harm. This specialized knowledge is precisely what the CN model, which is skewed towards Western epistemology, lacks (Adriana, Nadia 2025).

CN's consensus requirement conceals useful expert-backed information. X's focus on **"consensus among users who usually disagree"** rather than factuality is a "false premise of equaling truth with consensus". As a result, **over 85% of high-quality notes proposed that cite independent fact-checking organizations are still not visible** (Maldita 2025).

The two approaches can be made complementary, instead of being contradictory. The work of professional fact-checkers and community involvement are **"essential and complementary"** (Maldita 2025). Meta should avoid repeating the mistakes of X and should **collaborate with professional fact-checking organizations**. Solutions should involve **a close collaboration between experts, systems, and non-experts** to maintain quality and accountability.

Conclusion

Community Notes can play an important role in India, but it cannot, by design, address the country's linguistic breadth, polarized discourse, rapid misinformation spread, or harm-based narratives on its own. This is why

third-party fact-checking must coexist as an additional layer, especially since evidence suggests that CN performs better when informed by expert fact-checkers.

Examples of Community Notes

1. Automated content moderation systems can miss out on context and nuance

The Tweet below is written in Hindi language by the president of the Delhi unit of Aam Aadmi Party (AAP), a political party in India.

The tweet says, “11.11.2025 Chief Minister (CM) Rekha Gupta met the patient first. The next day on 12.11.2025, the Prime Minister (PM) met the same patient. New costume. New Plaster.”

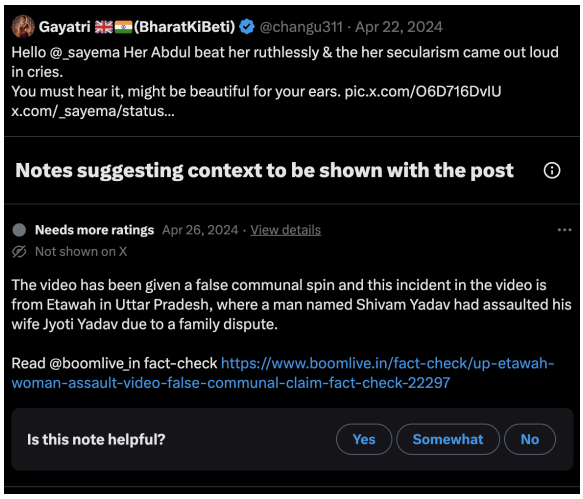
The tweet shares two images in which CM Rekha Gupta is seen interacting with a patient in a hospital on the left and the image on the right showing PM Narendra Modi meeting a patient on another day. The tweet was published on November 11, 2025, a day after a car exploded near Red Fort in Delhi killing several civilians. The views on the tweet shows its virality.

While the tweet does not make any direct claims, it insinuates that when the Prime Minister visited the patient he got a plaster while when the Chief Minister had visited he did not have a plaster, alluding that he is faking his injuries. In the comments section, we see that several users have posted a timeline of the Delhi CM meeting the patient several hours after the blast and the PM meeting the patient a day later explaining that the state of the patient, including his clothes, his appearance etc. are bound to be different and it cannot be used as evidence that the patient is faking his injuries. The tweet has no community note till date but there are fact-checks on the internet that have fact-checked the post verifying the timeline of events.

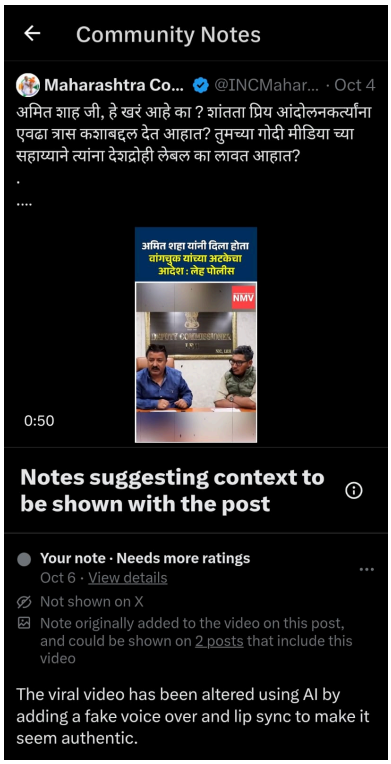
This tweet is an example of why context and nuance matter in combating misleading narratives that do not include a direct claim. The reason why such tweets are able to spread disinformation is because there have been instances in the past where images or associated text featuring political personalities meeting civilians in a crisis etc. have been debunked as manipulated. This is the kind of nuance and context that fact-checkers understand because they are part of a system that builds institutional knowledge. <tweet link>

2. Polarising tweet targeting a demographic doesn’t feature CNs as notes remain stuck waiting for cross-idealogical ratings

The claim made in the tweet below has been debunked by a fact-check article which has also been submitted in the note by a contributor, however, the note remains unpublished. This shows how cross-idealogical voting does not increase or ensure neutrality in information sharing on online spaces. <Tweet Link>



3. A tweet in Marathi consisted of an AI-manipulated video that had notes but didn't pass the rating threshold. <tweet link>



End Of Report