LEGAL NETWORK FOR TRUTHFUL ELECTIONS
Ground Floor, Jenkinsen Tower, 80 Timog Avenue, Diliman Quezon City
2nd Floor, Yap Sy Street cor. Andres Alonzo Street, Cotabato City *(beside Petron)*
lente.philippines@gmail.com
(02) 7502.1591 | 0917.106.6265 | 0947.164.4158 (Manila Office)
064 421 6007 (Cotabato Office)

Comments: Philippine Case on Political Misinformation and Third-Party Fact-Checking

1) Recent Trends on the Spread of Political Misinformation in the Philippines

Misinformation campaigns in the Philippines are an evolving problem. It remains as a key and prevalent issue, especially during election periods.

- **Evolution in strategies and tactics**
  - 2016: Characterized by the use of trolls and bots.
  - 2019: Saw the rise of "disinformation innovations" such as micro/macro-influencers and hyper-partisan “media” who did not have to hide behind anonymity.
  - 2022: Disinformation was rampant on "affiliate pages",[1] usually with less than 1000 followers, rather than on official candidate pages. Because of this, disinformation posts fly under the radar of fact-checkers, posing a challenge for monitoring.
  - 2025: The 2025 Philippine elections saw the same tactics and narratives employed as seen in previous elections. These include amplifying existing narratives of rigged elections and COMELEC corruption, flooding the information space, and harassment. Although fraud-related discussions account for only a small portion (2.1%) of overall election discussions,[2] they still carry significant weight in how people perceive the integrity of elections.
- **Political rivalries as the main triggers**
  The political rivalry between the Marcos and Duterte camps emerged as a key driver of disinformation during the election period. While both families have previously been associated with the use of disinformation to advance political

[1] Based on LENTE’s Social Media Monitoring Project for the 2022 National and Local Elections. Social Media Monitoring Tool used was Crowdtangle.
[2] Based on LENTE’s Social Media Monitoring in the 2025 National and Local Elections. Social Media Monitoring and Analysis Tools used include Brand24 and DisarmDecoderBot.

objectives, such tactics were particularly evident between February and March, when coordinated campaigns were observed targeting critics and discrediting rivals. This period saw heightened activity aimed at shaping public narratives and deepening political divides.

- **Government and Civil Society Responses**
    - **2019 Elections**: The policy against online disinformation began as a simple provision within the Implementing Rules and Regulations (IRR) of the Fair Elections Act (COMELEC Resolution 10488, Section 9.c). This acknowledged the emerging role of online platforms in online campaigning.
    - **2022 Elections**: The provision was slightly expanded (COMELEC Resolution 10730, Section 9c).
    - **2025 Elections:** COMELEC issued a separate, dedicated resolution (COMELEC Resolution 11064). This resolution includes provisions for mandatory registration of all official social media accounts and pages, AI-disclosure requirements, and prohibitions against the misuse of digital campaign platforms for the "purpose of disrupting or obstructing the election process or causing confusion among the voters…"
    - These policies, however, are limited by the existing legal framework on campaigning, which may not be adequate in capturing the nuances of the disinformation landscape.

2) The scope, adequacy, and impact of Meta's policies on political misinformation and Comments on the Third-Party Fact-Checking Program

a. Categories of Misinformation that are Removed

- **Meta's Policy:** Meta removes misinformation only in specific categories: Physical Harm or Violence, Harmful Health Misinformation, and Voter or Census Interference.
- **Potential Gap:** Meta currently removes misinformation only within narrowly defined categories: Physical Harm or Violence, Harmful Health Misinformation, and Voter or Census Interference. However, a significant gap remains in the platform's treatment of political misinformation that falls outside these categories. The video in question, for instance, does not relate to voting procedures or contain direct incitement to violence, placing it outside Meta's takedown thresholds. Instead, such content is typically subject to reduced distribution — a policy that may be insufficient in contexts like the Philippines, where disinformation has increasingly eroded public trust in democratic institutions.

  A further complication is that not all misleading political content is shared with malicious intent. In many cases, it is disseminated in the guise of expressing support for prominent figures, such as the former President, or as part of identity-based political discourse. Nonetheless, even unintentional or emotionally driven sharing of false narratives can contribute to voter confusion, deepen political polarization, and distort public debate, particularly during sensitive electoral periods.

- **Recommendations:**
  - Meta should establish and publish clear criteria for identifying highly polarized electoral contexts. These may include histories of political violence, documented use of disinformation in past elections, or intensified online harassment of political actors. This would enable Meta to adopt a more proactive and responsive moderation approach during critical periods.
  - Meta should consider broadening its misinformation policies to include content that undermines democratic processes, even when it does not meet the threshold for direct voter interference. This includes politically charged disinformation or misleading narratives about public institutions, candidates, or electoral processes.
  - Immediate and visible labels should be applied to content flagged as potentially false or misleading especially those involving civic or political claims while fact-checking is ongoing. This would help limit the viral spread of harmful content during high-risk periods.

- Meta must tailor enforcement strategies to reflect local political contexts. In the Philippines, for example, the emotional appeal of political loyalty can drive the organic spread of disinformation. Local civil society organizations and fact-checkers should be engaged to help detect such patterns early and design more culturally attuned interventions.

b. Prioritization for Human Review

- **Meta's Policy:** Meta's policy for human review prioritizes content based on three factors: **Severity** (potential for harm), **Virality** (speed of sharing), and **Likelihood of Violating** its policies. The company states that it prioritizes "high-severity content with the potential for offline harm and viral content which is spreading quickly."
- **Potential Gaps:**
  - In the present case, a Meta classifier identified the content as "possible misinformation" and it was reported by a user, but it was not acted upon, possibly because the criteria, "Likelihood of Violating" and "Virality" were not met, nor it did not fall under the most severe categories like "imminent physical harm".
  - A post that is algorithmically identified as "possible misinformation" but still isn't prioritized for review is a bottleneck in the existing system. This may be understandable given the massive scale of disinformation campaigns that overwhelm existing fact-checking capacity.
- **Recommendations:**
  - As Meta implements alternative content moderation models like the Community Notes program in the US and explores implementing it in other jurisdictions like the Philippines, the case emphasizes the need to ensure that a responsive system is in place by maximizing not only traditional fact-checking groups.
  - Partnerships beyond fact-checking organizations to include media literacy organizations, academic institutions, human rights organizations, and other grassroots organizations may be maximized as contributors for the community notes program.
  - Meta can explore revising its prioritization guidelines to ensure that content flagged by its classifiers as "likely to be misinformation" is subjected to a mandatory human review, regardless of whether it meets the highest virality or severity thresholds.