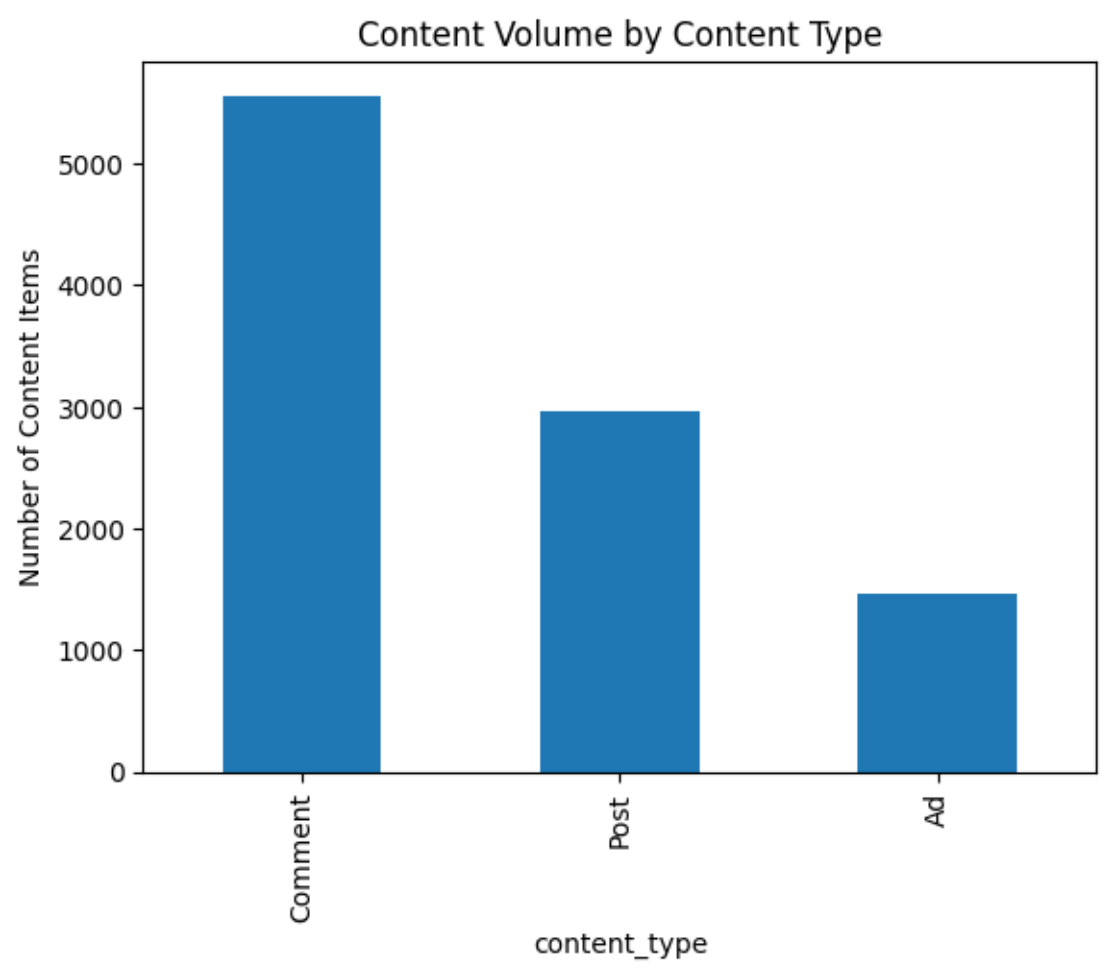
Diagnosing Contextual Content Moderation Gaps in the India–Bengali Market

Contextual offensive content in the India–Bengali market is systematically under-detected, leading to delayed enforcement and silent harm

- AI moderation avoids over-flagging benign content but **misses a significant share of offensive posts**
- Missed content remains live **~2× longer**, increasing user harm
- Risk concentrates in **comments, crisis periods, and contextual content**
- **Market-specific operational interventions** are required to close the gap

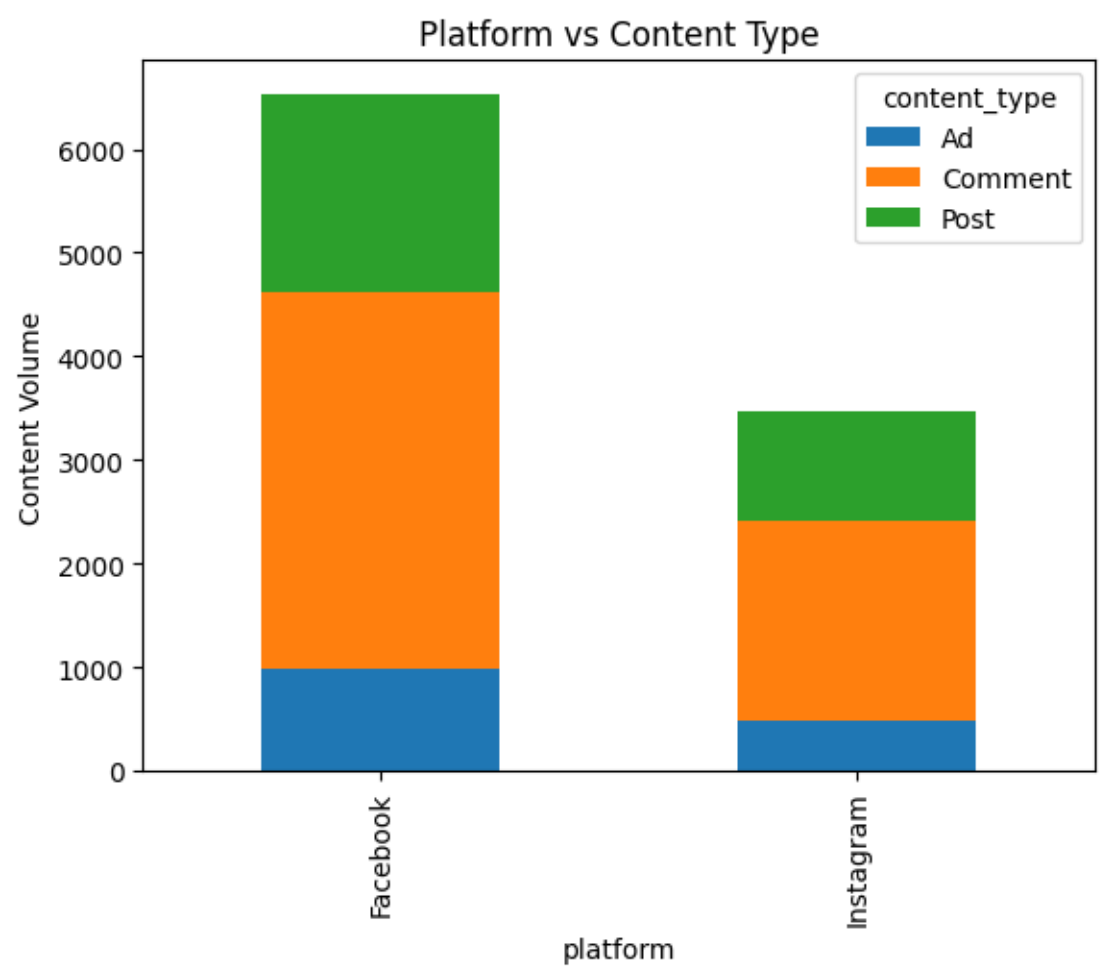
High-volume, user-generated content dominates the moderation workload

- **Comments ≈ 50% of total content volume**
- Posts (~30%) and Ads (~15%) form smaller shares
- User interactions, not ads, drive moderation exposure

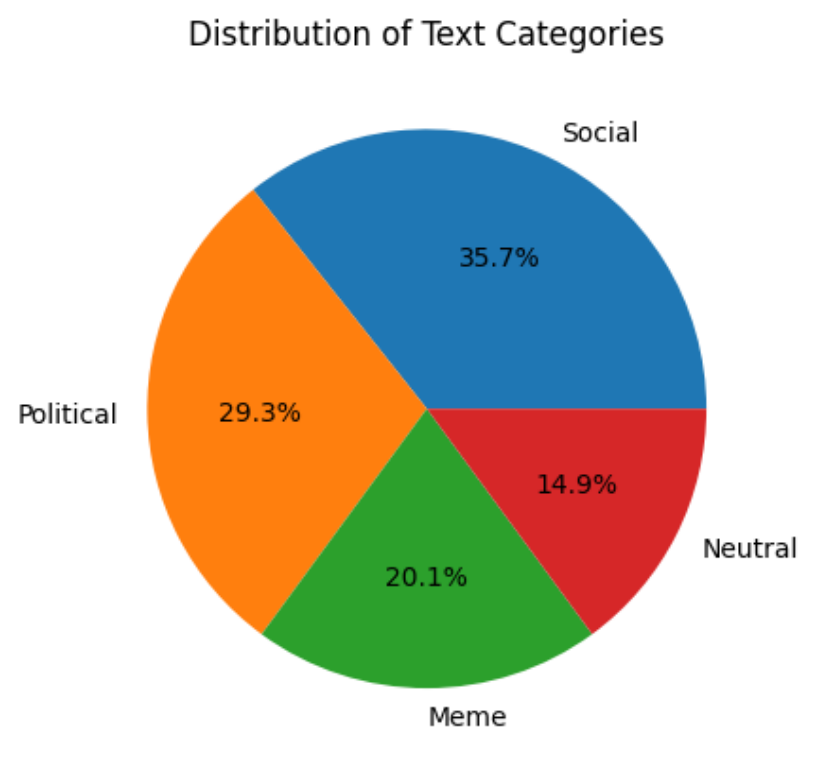


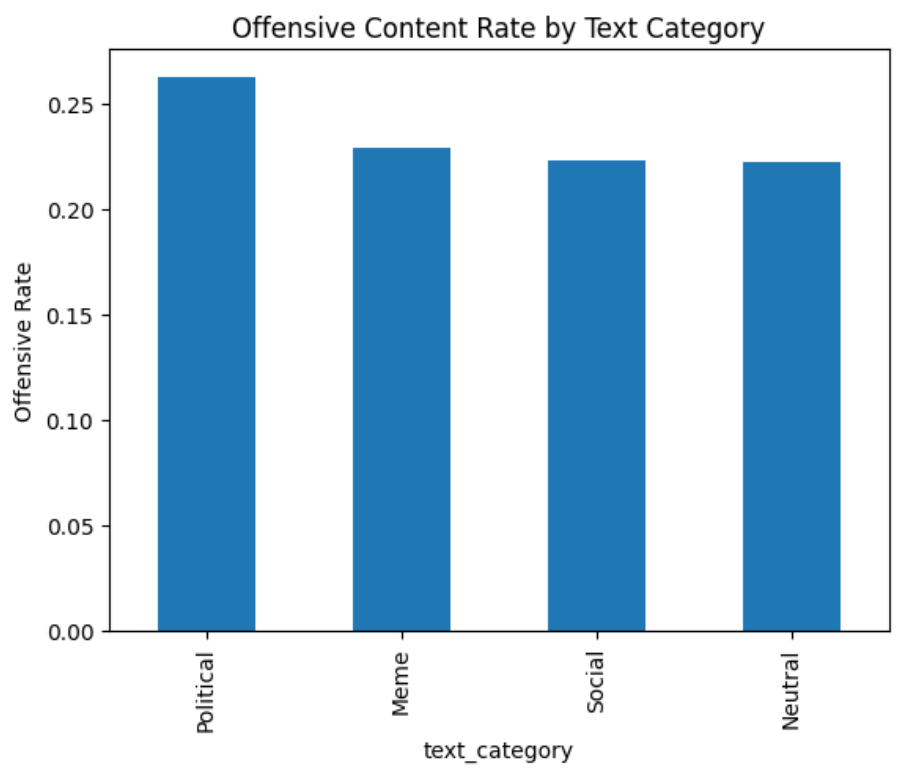
Facebook carries higher and more diverse moderation risk than Instagram

- **Facebook** dominates overall volume and content diversity
- Instagram is **comment-heavy**, amplifying conversational risk
- Platform mix affects how enforcement capacity must be allocated



Most content is socially or politically expressive, limiting keyword-based detection





Nearly **50%** of content is politically or meme-tically expressive

These categories rely heavily on context, tone, and implication, not explicit abuse

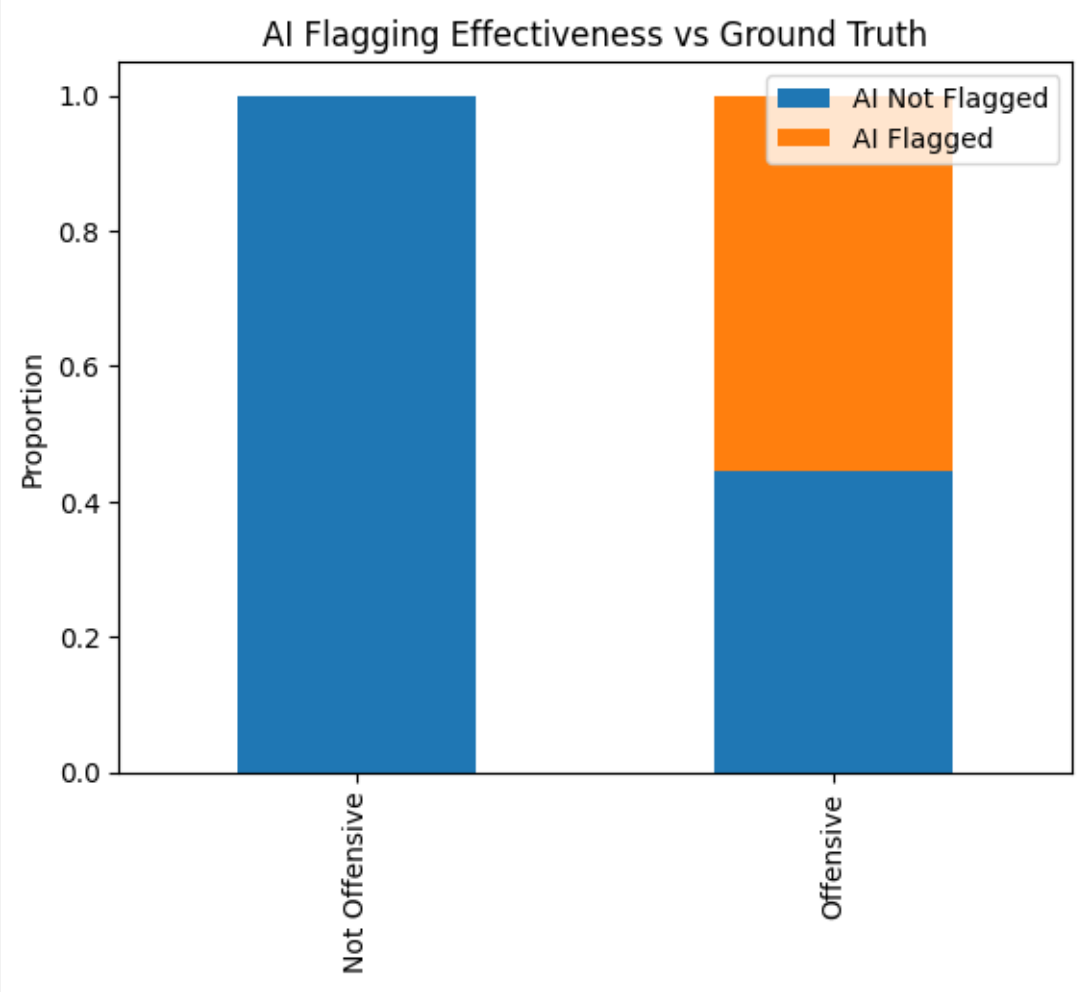
Offensiveness is **not confined to "expected" categories** — even neutral-looking content carries risk.

AI moderation avoids false positives but fails to catch a large share of offensive content.

- For non-offensive content, AI correctly leaves ~90% unflagged
- For offensive content, AI misses ~40–45%

Indicates high precision, low recall

The system is conservative by design – but this creates silent exposure risk.



Detection failures intensify at specific category × crisis intersections.

Highest false-negative risks:

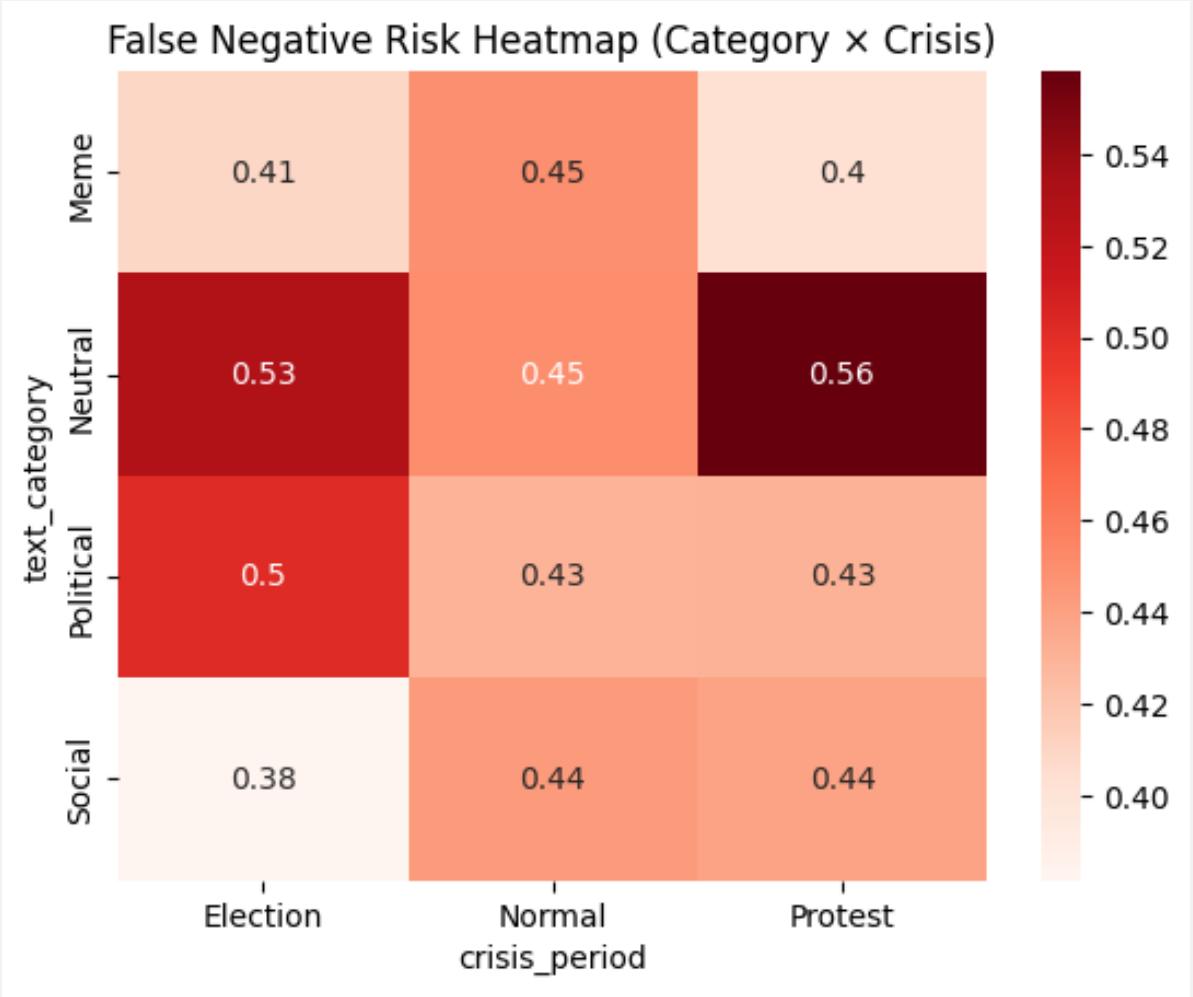
- Neutral × Protest (~0.56)
- Meme × Protest (~0.56)
- Neutral × Election (~0.53)

Lowest risk:

- Social × Election (~0.38)

Seemingly benign or humorous content becomes dangerous in volatile contexts

AI struggles most when context flips meaning



AI misses significantly increase time-to-action, extending user harm.

No false negatives:

- Median TTA ≈ 20 hours
- Tight distribution, low variance

False negatives:

- Median TTA ≈ 40 hours
- Long tail up to **120 hours**

Missed content stays live 2× longer, compounding harm during sensitive periods.

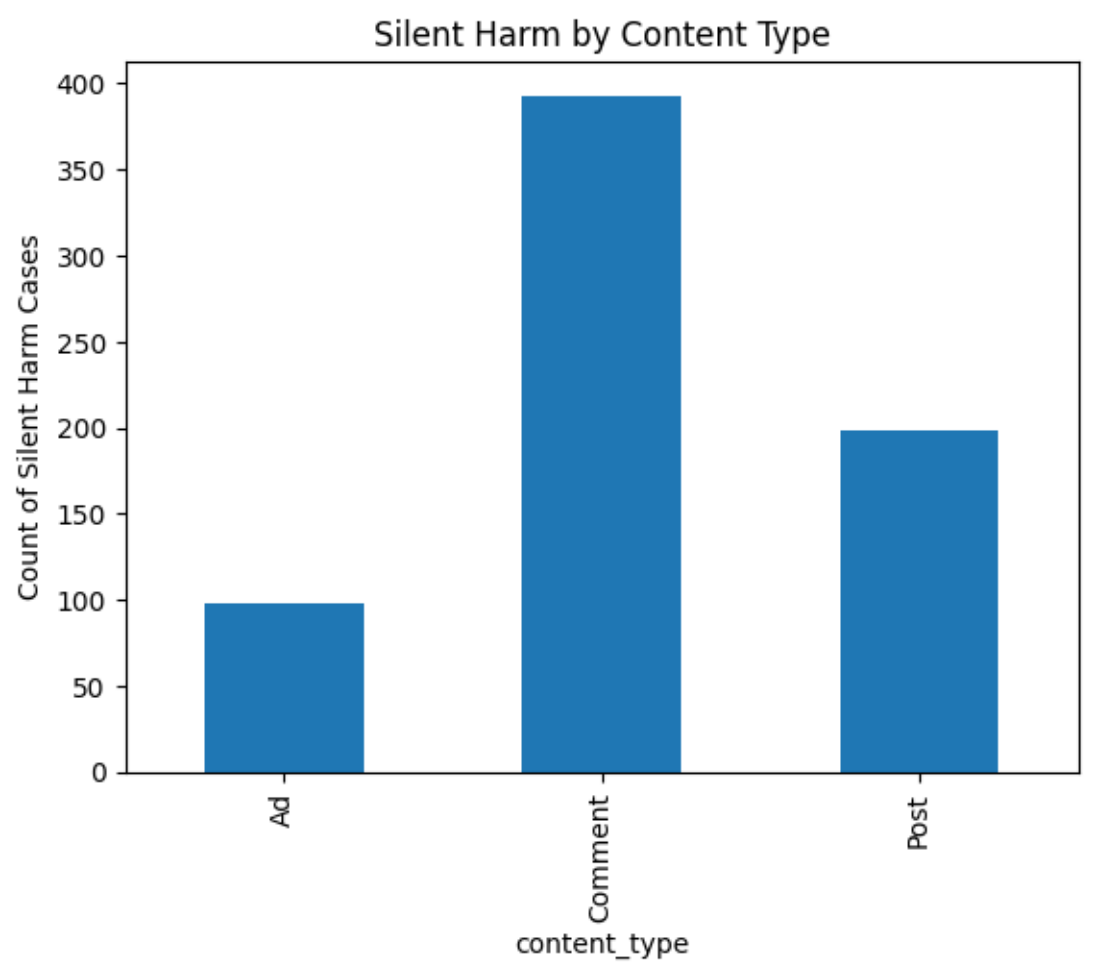
High-volume comment streams are the largest source of undetected harm.

Silent harm cases:

- Comments: ~350
- Posts: ~200
- Ads: ~100

Scale + conversational dynamics make comments hardest to police

Reliance on user reporting is insufficient



Resolution: Market-specific, context-aware operations can materially reduce risk without over-enforcement.

Contextual Risk Routing

Prioritize human review for:

- Comments
- Meme & Neutral content
- Protest & Election periods

False-Negative Monitoring

- Track category × crisis risk patterns as operational KPIs
- Use these signals to guide escalations

Crisis Playbooks

Expect higher false negatives during crises

Pre-emptively allocate review capacity

Expected Outcomes

- **Faster time-to-action**
- **Reduced silent harm**
- **Lower escalation risk**