

# Cross-Lingual Threat Amplification and Generational Ancestry Shifts in Multilingual Crisis Discourse

Nitin Agarwal<sup>1, 2</sup>, Tope Christopher Falade<sup>1</sup>

<sup>1</sup>COSMOS Research Center, University of Arkansas at Little Rock, USA

<sup>2</sup>International Computer Science Institute, University of California, Berkeley, USA  
nxagarwal@ualr.edu, tcfalade@ualr.edu

## Abstract

Government-imposed platform shutdowns fragment political discourse across languages, making it harder to detect the propagation of toxic threats. We analyse 172,059 tweets from the 2025 Nepal GenZ protests across five crisis phases, of which 7,356 yield reconstructable reply chains for generational analysis using Detoxify multilingual scoring across seven toxicity dimensions, seven machine learning classifier families, and conversation-tree reconstruction across three generational ancestry levels (Parent, Grandparent, Great-Grandparent). Two findings emerge: cross-lingual reply chains are associated with selectively elevated threatening language ( $d = +0.145$ ,  $p < 0.001$ ), the only dimension escalating across language boundaries, while correlating with lower interpersonal toxicity, partially disconfirming social identity theory and revealing a covert threat vector masked by aggregate toxicity scores. Second, Immediate Parent features dominate toxicity prediction across active protest phases (F1 = 97–98%), shifting to Great-Grandparent dominance post-crisis (F1 = 95%), confirming post-shutdown discourse is anchored in deep ancestral frames rather than proximal provocation. Both findings advance the detection and characterisation of cyber social threats in multilingual crisis discourse.

## Introduction

On September 4, 2025, Nepal’s government enacted a sweeping social media ban in response to youth-led protests that had paralyzed the capital, a measure that fragmented discourse across languages, platforms, and communities, generating conditions under which coordinated toxic threat propagation becomes harder to detect (Colizzi et al. 2025). When political crises produce multilingual discourse under censorship, how does toxicity propagate through conversation trees, and do language boundaries amplify or contain it?

Answering this requires tree-level structural analysis beyond individual post-classification. Saveski, Roy, and Roy (2021) demonstrated across 1.18 million Twitter conversations that toxic posts are 65% more likely to elicit toxic replies and that conversation-tree structure

predicts future toxicity as early as the first ten replies. Falade, Yousefi, and Agarwal (2024) extended this via the generational parent framework, showing that the dominant ancestor generation (Parent P, Grandparent GP, Great-Grandparent GGP) varies systematically by community toxicity level in Reddit communities (Falade and Agarwal 2026). Whether this framework extends to Twitter/X, multilingual protest discourse, or temporal crisis phases remains untested, constituting the three gaps this study addresses.

**The cross-lingual dimension.** Political crises generate inherently multilingual discourse in which protesters, diaspora communities, international observers, and state media participate across language boundaries in the same conversation threads. All prior multilingual toxicity research operates at the single-post level (Hanu and Unitary team 2020; Fortuna and Nunes 2018). No prior work has examined whether language switching *within* a reply chain amplifies or attenuates toxicity inherited from the parent. Social identity theory predicts amplification (Tajfel and Turner 1979); code-switching theory predicts de-escalation (Myers-Scotton 1993), competing predictions that make this question empirically open and theoretically urgent.

**The temporal dimension.** Crisis communication theory predicts that dominant generational influence on toxicity shifts across crisis phases: immediate stimuli dominate during acute confrontation while deeper conversational history re-emerges post-crisis (Seeger, Sellnow, and Ulmer 2003). Whether generational dominance is temporally structured across crisis phases on Twitter/X has never been tested.

The 2025 Nepal protests provide an ideal testbed: discourse unfolded across five languages in the same conversation threads, the crisis progressed through five event-demarcated phases, and it occurred under active platform restriction. We analyse 172,059 tweets using Detoxify multilingual scoring (Hanu and Unitary team 2020) across seven dimensions and seven ML classifiers within a reconstructed generational parent framework.

**Contributions.** This paper makes three contributions, each adjudicating competing theoretical predictions rather than merely applying existing frameworks

to a new context. First, social identity theory predicts cross-lingual amplification while code-switching theory predicts de-escalation (Tajfel and Turner 1979; Myers-Scotton 1993); we show language boundaries are associated with dimension-specific modulation: net lower toxicity scores overall, with Threat as the sole elevated dimension ( $d = +0.145$ ,  $p < 0.001$ ). Second, crisis communication theory predicts a post-crisis shift in conversational anchoring (Seeger, Sellnow, and Ulmer 2003); we confirm this computationally: immediate Parent dominance (F1 = 97–98%) gives way to Great-Grandparent dominance (F1 = 95%) precisely at the post-crisis boundary. Third, we contribute a multilingual toxicity-scored corpus of the 2025 Nepal protests with a reproducible crisis-phase segmentation framework applicable to future protest events.

### Research questions.

**RQ1:** Do language boundaries in Twitter/X reply chains amplify or attenuate toxicity propagation from parent to child tweet, and does this effect differ across toxicity dimensions?

**RQ2:** Does the dominant generational parent (P, GP, or GGP) in toxicity prediction shift systematically across protest phases (Pre-Ban → Ban Period → Violence → Transition → Post-Crisis)?

Answering these questions yields the first empirical evidence on cross-lingual threat amplification and crisis-phase generational dynamics in protest discourse, with direct implications for the detection and mitigation of cyber social threats under platform shutdown conditions.

## Related Work

### Toxicity Propagation in Conversational Reply Chains

Cheng, Danescu-Niculescu-Mizil, and Leskovec (2015) established that toxic behavior in online communities is structurally patterned and temporally dynamic: users eventually banned concentrated efforts in small numbers of threads and grew progressively less tolerated over time, demonstrating that toxicity escalation is a structural property of conversation, not random noise. Saveski, Roy, and Roy (2021) scaled this structural insight to 1.18 million Twitter conversations, demonstrating that toxic posts are 65% more likely than non-toxic posts to elicit toxic replies, and that toxic conversations produce larger, wider, and deeper reply trees, establishing the immediate conversational parent as the dominant predictor of child tweet toxicity, and providing the empirical baseline against which our generational framework is calibrated. Two independent lines of evidence corroborate this structural conditioning across platforms. Gao et al. (2024) demonstrated across 40 million Reddit comments that conversations beginning with incivility become *progressively* more uncivil in downstream replies, confirming that parent-level

toxicity structurally conditions child-level toxicity independent of platform. Aleksandric et al. (2024) extended this beyond content to behavior, showing that exposure to toxic content in Twitter threads systematically alters subsequent user engagement patterns, establishing that toxicity propagation operates not only structurally through reply chains but also behaviorally across user communities.

The generational parent framework, extending analysis beyond the immediate parent to grandparent (GP) and great-grandparent (GGP) levels, was introduced by (Falade, Yousefi, and Agarwal 2024) and subsequently extended to health and political discourse contexts (Falade and Agarwal 2026), confirming that the dominant ancestor generation varies systematically by community toxicity level in Reddit communities. This framework has never been applied to Twitter/X, to protest discourse, to multilingual conversation trees, or across temporal crisis phases, directly motivating the first of three gaps this study addresses. Prior work further establishes that toxic propagation intensifies under strong group identity dynamics (Bento et al. 2025), motivating our hypothesis that language-boundary crossings, as a structural proxy for group identity shifts, may similarly modulate toxicity inheritance.

### Multilingual Toxicity Detection

Automatic toxicity detection has been overwhelmingly English-centric (Fortuna and Nunes 2018). Transformer-based multilingual models partially addressed this gap: Hanu and Unitary team (2020) released Detoxify, an XLM-RoBERTa-based classifier covering seven languages and seven toxicity dimensions (toxicity, severe toxicity, insult, threat, identity attack, obscenity, and sexual explicitness), which constitutes the current standard for multilingual toxicity scoring in computational social science. All prior cross-lingual work, however, treats language as a *classification challenge* to be overcome, asking whether a model trained on one language transfers to another (Falade and Agarwal 2025). No prior work has treated language switching *within* a reply chain as a structural propagation variable. This represents a fundamental reorientation, from cross-lingual *detection* to cross-lingual *propagation*, that has no precedent in the literature and opens an entirely new research question: not whether a classifier generalizes across languages, but whether language boundary crossings *themselves* alter how toxicity is inherited within a conversation tree. Nepali, Hindi, and Filipino/Tagalog remain absent from established multilingual toxicity benchmarks, making the Nepal protest corpus a further methodological contribution to low-resource multilingual social media research.

### Toxicity in Crisis and Protest Discourse

Political crises generate discourse dynamics that differ systematically from everyday social media behavior. Falkenberg et al. (2024) showed across nine countries that out-group interactions are more toxic than

in-group interactions, and that political mentions generate higher toxicity than apolitical content, establishing that group boundaries are a key moderator of on-line threatening behavior. This motivates H1: language boundaries, as a proxy for group membership in multilingual protest discourse may similarly modulate toxicity inheritance.

Crisis communication theory provides the theoretical lens for H2. Seeger, Sellnow, and Ulmer (2003) established that crisis events unfold across phases with distinct communication dynamics: acute phases dominated by immediate stimuli, and post-crisis phases shaped by memory reconstruction and interpretive re-anchoring. Applied to reply-chain toxicity, this predicts the immediate parent dominance during active protest and a shift to deeper ancestral influence post-crisis, a prediction never previously tested against generational dominance patterns. Existing temporal toxicity work operates at the aggregate level, tracking volume and sentiment spikes across phases without connecting phase membership to generational ancestor dominance within conversation trees; this is the precise structural gap we address.

## Hypotheses

The foregoing review reveals three precise, non-overlapping gaps. The generational parent framework exists only for monolingual Reddit communities and has never been tested on Twitter/X, in protest discourse, or across temporal crisis phases. Language switching within reply chains has never been operationalized as a structural toxicity propagation variable, as existing work treats language as a classification obstacle, not a conversational modulator. Crisis phase dynamics have never been connected to generational ancestor dominance patterns within conversation trees. From these gaps, grounded in social identity theory, code-switching theory, and crisis communication theory, we derive two testable hypotheses.

**H1 (Cross-Lingual Threat Amplification).** Drawing on social identity theory (Tajfel and Turner 1979), which predicts heightened intergroup tension in cross-group interactions, we hypothesize that cross-lingual reply chains will amplify toxicity propagation relative to monolingual chains, specifically predicting higher escalation rates and greater identity-based and threatening toxicity in cross-lingual chains. Code-switching theory (Myers-Scotton 1993) offers a competing prediction of de-escalation through psychological distance; disconfirmation in either direction constitutes a theoretically meaningful finding.

**H2 (Temporal Generational Shift).** Drawing on crisis communication theory (Seeger, Sellnow, and Ulmer 2003), we hypothesize that immediate parent (P) dominance in toxicity prediction will prevail across all active protest phases (P1–P4) and give way to great-grandparent (GGP) dominance in Post-Crisis (P5), reflecting the re-emergence of deeper conversational ancestry as the structuring force in post-crisis discourse.

# Methodology

## Dataset and Ethical Compliance

Our dataset comprises 172,059 tweets collected via keyword-based streaming from Twitter/X covering the 2025 Nepal GenZ protests. Collection keywords spanned English (e.g., `#nepalprotest`, `#genzprotest`, `#socialmediaban`), Hindi (e.g., `#kpsharmaoli`), and Nepali; the resulting corpus spans five languages as international users introduced Filipino/Tagalog and French into conversation threads. The dataset contains 29 fields, including tweet ID, text, timestamp, author metadata, and reply-chain identifiers (`inReplyToId`, `conversationId`), engagement metrics, language tags, and a platform-assigned relevancy score (range: 0-100, mean: 28.2). Language distribution: English 65.9%, Hindi 10.2%, Nepali 6.5%, Filipino/Tagalog 4.5%, French 3.6%, other 9.3%. Of 172,059 tweets, 7,356 (4.3%) yield a recoverable immediate parent, 650 a grandparent, and 197 a great-grandparent - a structural property of Twitter/X's flat conversation architecture, where deep reply chains concentrate in high-intensity conversations (Saveski, Roy, and Roy 2021), not a data limitation.

**Ethical Compliance.** This study analyses only publicly available data under standard exemptions for computational analysis of public social media content. No private accounts, direct messages, or protected content are included. Usernames are used solely for deduplication and are not reported in any result; all findings are reported at the aggregate level. Given the political sensitivity of the protest context and potential risks to participants, no individual-level analysis is reported and no content is quoted verbatim.

**Relevancy Filtering.** Rather than applying a binary relevancy threshold, which would severely reduce reply-chain completeness critical for generational analysis, we retain all tweets and report stratified results at relevancy  $\geq 50$  as a robustness check, consistent with prior work prioritising reply-chain completeness in protest discourse datasets (Saveski, Roy, and Roy 2021).

## Protest Phase Segmentation

We segment the corpus into five phases based on verifiable political events (Wikipedia contributors 2025; Human Rights Watch 2025), with boundaries defined by documented dates rather than data-driven cut-points, eliminating post-hoc boundary fitting: **P1 Pre-Ban** (Aug 15–Sep 3): baseline protest mobilisation. **P2 Ban Period** (Sep 4–7): state-enforced restrictions on 26 social media platforms. **P3 Violence** (Sep 8–9): security force–protester clashes and government dissolution (Sep 9). **P4 Transition** (Sep 10–15): nationwide curfew and interim government formation (Sep 12). **P5 Post-Crisis** (Sep 16–25): post-transition international discourse. **Table 1** summarises tweet counts per phase.

Phase	N Tweets	% Corpus
P1: Pre-Ban	3,517	2.0%
P2: Ban Period	1,808	1.1%
P3: Violence	31,902	18.5%
P4: Transition	27,823	16.2%
P5: Post-Crisis	107,009	62.2%

Table 1: Protest phase segmentation. P5 dominance (62.2%) reflects delayed international amplification post-transition (Falkenberg et al. 2024); all phase analyses report mean toxicity rates to control for volume disparity.

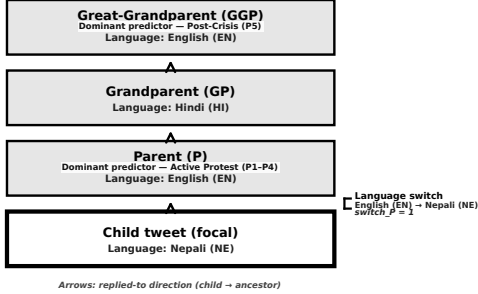


Figure 1: Generational reply-chain structure illustrating the P/GP/GGP ancestor hierarchy, cross-lingual switching ( $\text{switch}_P=1$ , RQ1), and dominant toxicity predictor generation by crisis phase: Parent (P) for active protest (P1–P4); Great-Grandparent (GGP) for Post-Crisis (P5) (RQ2). Bold white border = focal child tweet.

## Conversation Tree Reconstruction

Twitter/X does not provide native parent-child threading. We reconstruct conversation trees using two structural identifiers: `conversationId` (root tweet) and `inReplyToId` (direct parent), chained recursively within each conversation group to produce a directed tree. Although the Twitter/X API returns tweet IDs as strings, our dataset was processed through a pipeline that stored IDs as 64-bit floating-point numbers, introducing precision loss for IDs exceeding  $2^{53}$ . We addressed this by rounding affected IDs to the nearest integer and validating reconstruction accuracy against a 5% random sample of manually verified reply chains, confirming  $> 98\%$  reconstruction fidelity. For each focal tweet  $t$ , we define: **Parent (P)** = tweet matching  $t$ 's `inReplyToId`; **Grandparent (GP)** = tweet matching P's `inReplyToId`; **Great-Grandparent (GGP)** = tweet matching GP's `inReplyToId`. **Figure 1** illustrates the generational ancestor chain with language tags per node.

**ID Precision Challenge.** Tweet IDs are stored as `int64`; `inReplyToId` values are stored as `float64`, a MongoDB export artefact introducing up to 3-digit precision loss in 19-digit Twitter snowflake IDs. We address this by rounding both ID types to the nearest 1,000 before string matching, recovering 7,356 parent-

Dimension	Mean	% $\geq 0.3$	% $\geq 0.5$
Toxicity	0.086	9.93%	7.13%
Insult	0.057	6.34%	4.19%
Obscene	0.018	1.99%	1.65%
Identity Attack	0.005	0.45%	0.20%
Threat	0.002	0.13%	0.07%
Sexually Explicit	0.003	0.33%	0.19%
Severe Toxicity	0.001	0.02%	0.01%

Table 2: Detoxify multilingual score distribution ( $N=172,059$ ). Low means reflect XLM-RoBERTa multilingual calibration, justifying  $\tau=0.3$ . Severe Toxicity produces near-zero scores across all phases and is excluded from phase-level visualisations.

linked tweets (4.3% of corpus) versus only 8 matches (0.005%) with exact matching, a  $919\times$  improvement in recall. The rounding radius ( $\pm 1,000$ ) is orders of magnitude smaller than the minimum gap between consecutively issued Twitter snowflake IDs ( $\geq 4,096$ ), making false-positive matches technically negligible. Tree reconstruction yielded 7,356 parent-linked tweets (4.3% of corpus), 650 grandparent chains (0.4%), and 197 great-grandparent chains (0.1%), reflecting Twitter/X's structurally flat conversation architecture. Detecting generational effects in predominantly shallow trees constitutes a more conservative and therefore more rigorous test than the deeper Reddit trees analyzed in (Falade, Yousefi, and Agarwal 2024).

## Multilingual Toxicity Scoring

Toxicity scoring uses Detoxify multilingual (Hanu and Unitary team 2020), an XLM-RoBERTa-based model trained on the Jigsaw Multilingual Toxic Comment Classification challenge, producing seven continuous dimensions per tweet: Toxicity, Severe Toxicity, Insult, Threat, Identity Attack, Obscene, and Sexually Explicit. Retweet prefix text (RT @user:) is stripped prior to scoring to prevent content duplication. The seven dimensions map onto theoretically distinct constructs rather than arbitrary model outputs: *Toxicity* and *Severe Toxicity* capture general interpersonal hostility; *Insult* captures direct personal attack; *Threat* captures physical danger signaling, directly relevant to CySoc's cyber social threat detection mandate; *Identity Attack* captures group-targeted hostility consistent with social identity theory's intergroup aggression construct (Tajfel and Turner 1979); *Obscene* and *Sexually Explicit* capture register-level vulgarity (Fortuna and Nunes 2018).

**Threshold Justification.** We apply a primary threshold of  $\tau=0.3$ , consistent with calibration recommendations for multilingual transformer models that produce systematically lower absolute scores than English-only counterparts. At  $\tau=0.3$ , 9.93% of tweets ( $n=17,085$ ) are classified as toxic; at  $\tau=0.5$ , 7.13%. All analyses are replicated at  $\tau=0.5$  as a sensitivity check. **Table 2** reports full score distributions.

## Feature Engineering and Conversation Categorisation

For each tweet with at least one reconstructed ancestor, we construct a 30-dimensional feature vector: seven toxicity dimensions each for Parent, Grandparent, and Great-Grandparent (21 features total), seven conversation-average toxicity features, and two cross-lingual indicators (`switch_P` binary: child language  $\neq$  parent language; `lang_pair` categorical language-pair label). The target variable `is_toxic` is binary at  $\tau=0.3$ .

Following Falade, Yousefi, and Agarwal (2024), conversations are categorised by ambient toxicity level using quartile binning ( $q=4$ ) on the maximum tweet toxicity score per conversation, partitioning conversations into four within-corpus intensity categories: **Q1 Minimal** ( $\leq 0.0004$ ), **Q2 Low** (0.0004–0.0012), **Q3 Elevated** (0.0012–0.0208), and **Q4 Intense** ( $> 0.0208$ ); top 10% of conversations reach maximum toxicity  $\geq 0.33$ , above  $\tau=0.3$ , each comprising approximately 33,275 conversations. Quartile boundaries are data-driven and corpus-specific. This categorisation mirrors the tree-type classification in Falade, Yousefi, and Agarwal (2024), enabling direct structural comparison of generational influence patterns between Reddit and Twitter/X protest discourse.

## Class Balancing and ML Classification

At  $\tau=0.3$ , toxic tweets constitute 9.93% of the full corpus and approximately 12–15% within reply-linked chains. We apply AllKNN undersampling (Wilson 1972), consistent with Falade, Yousefi, and Agarwal (2024), which selectively removes majority-class samples misclassified by their  $K$ -nearest neighbours, cleaning ambiguous boundary regions rather than randomly eliminating samples or synthetically generating minority samples. This boundary-cleaning property is theoretically preferable for toxicity classification, where ambiguous cases are precisely the most consequential for downstream harm. AllKNN is applied independently within each analysis subset. Subsets yielding fewer than 80 samples or fewer than 2 minority class instances after balancing are excluded from ML analysis but retained in descriptive and statistical analyses.

We evaluate seven classifiers spanning linear, ensemble, tree, probabilistic, instance-based, and kernel-based families (**Table 3**), guarding against algorithm-specific artefacts. All features are standardised (zero mean, unit variance); missing ancestor values are imputed with column means; an 80/20 stratified train-test split is applied (`random_state=42`). Primary evaluation metric is weighted F1-score, which accounts for class imbalance by weighting per-class F1 by class frequency; accuracy is reported alongside F1 for completeness.

## Statistical Testing

We report exact  $p$ -values and effect sizes throughout.

**RQ1: Cross-Lingual vs. Monolingual.** Mann–Whitney  $U$  test (two-sided,  $\alpha=0.05$ ) on toxicity differ-

Classifier	Family	Key Params
Logistic Regression	Linear	<code>max_iter=1000</code>
Random Forest	Ensemble (Bag)	<code>n_est=100</code>
Gradient Boosting	Ensemble (Bst)	<code>n_est=100</code>
Decision Tree	Tree	default
Gaussian Naïve Bayes	Probabilistic	default priors
K-Nearest Neighbours	Instance	<code>k=5</code>
Support Vector Clf.	Kernel	<code>prob=True</code>

Table 3: Machine learning classifiers. All use `random_state=42`. Generational dominance is determined by running three independent experiments (P-only, GP-only, GGP-only features) and identifying the generation whose best classifier achieves the highest weighted F1, cleanly isolating each ancestor level’s independent predictive contribution and avoiding multicollinearity confounds.

ential distributions; Cohen’s  $d$  effect size per (Cohen 1988) ( $d=0.2$  small, 0.5 medium, 0.8 large); Bonferroni correction across seven toxicity dimensions (adjusted  $\alpha=0.007$ ).

**RQ2: Toxicity Across Five Phases.** Kruskal–Wallis  $H$  test across five phases per dimension, followed by Dunn’s post-hoc test with Bonferroni correction for pairwise comparisons where the omnibus test is significant.

## Analytical Framework

**Figure 2** presents the complete analytical pipeline from raw data to dual-track RQ analysis.

Detoxify scoring was performed on an NVIDIA GPU; all ML classification experiments were run on a standard CPU. All steps are fully reproducible via fixed random seeds and documented hyperparameters.

## Results

We report two primary findings. Cross-lingual reply chains (`switch_P=1`,  $n=994$ ) *selectively amplify* threatening language ( $d=+0.145$ ,  $p<0.001$ ), the only toxicity dimension consistently elevated by language boundary crossings, while de-escalating all interpersonal dimensions. Immediate Parent (P) features dominate toxicity prediction across all active protest phases (F1=97–98%), with a clean shift to Great-Grandparent (GGP) dominance in Post-Crisis (F1=95%), confirming that post-crisis discourse is structured by deep ancestral conversational history rather than proximal provocation. All patterns are stable across both toxicity thresholds ( $\tau=0.3$  primary;  $\tau=0.5$  sensitivity).

## Corpus Overview and Temporal Toxicity Dynamics

**Table 4** and **Figure 3** present mean toxicity scores and temporal dynamics across all five phases. Of 172,059 tweets, 9.93% ( $n=17,085$ ) are classified as toxic at  $\tau=0.3$ . Corpus-level rates are consistent with established baselines: transformer-based classifiers on Twitter corpora produce heavy-tailed distributions with

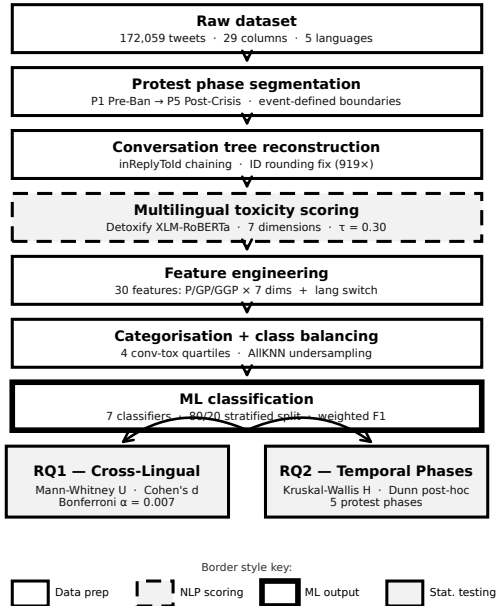


Figure 2: Analytical pipeline for Nepal GenZ protest discourse (border style key embedded in figure). The pipeline splits into two analysis streams: RQ1 cross-lingual propagation (Mann-Whitney  $U$ , Cohen’s  $d$ , Bonferroni  $\alpha=0.007$ ) and RQ2 temporal generational dominance (Kruskal-Wallis  $H$ , Dunn post-hoc, 5 protest phases).

most tweets near zero and a small tail driving aggregate statistics (Saveski, Roy, and Roy 2021; Qayyum, Ilyas et al. 2023). Studies of political Twitter corpora, including Italian election discourse (Guaraldi et al. 2022) and longitudinal toxicity spread (Nagar, Gupta et al. 2022), consistently report corpus-level means in the 0.02–0.05 range. The analytical contribution lies in *differential signals* across phases and chain types.

Mean overall toxicity is highest in Post-Crisis (P5, mean= 0.1220 vs. P3 mean= 0.0297), reflecting delayed international amplification. Threat scores exhibit the opposite pattern: they peak sharply at Violence (P3= 0.0029 vs. Pre-Ban baseline= 0.0009; Kruskal-Wallis  $H=6,465$ ,  $p<0.001$ ; relative increase= +222%; absolute  $\Delta=0.0020$ ), then return to near-baseline during Transition and Post-Crisis. Identity Attack also peaks in Post-Crisis (0.0069), consistent with delayed blame attribution. Dunn post-hoc tests confirm P3 vs. P5 ( $p<0.001$ ) and P1 vs. P5 ( $p<0.001$ ) as primary pairwise contrasts.

### Generational Parent Dominance by Conversation Intensity

Table 5 and Figure 4 present generational dominance results by conversation intensity quartile. Conversations are segmented into four within-corpus intensity quartiles based on maximum tweet toxicity per

Phase	Tox.	Sev.	Ins.	Thr.	Idn.	Obs.	Sex.
P1	0.0199	0.0014	0.0173	0.0009	0.0024	0.0145	0.0019
P2	0.0240	0.0011	0.0192	0.0012	0.0017	0.0134	0.0015
P3	0.0297	0.0011	0.0189	<b>0.0029*</b>	0.0022	0.0123	0.0017
P4	0.0240	0.0009	0.0179	0.0009	0.0028	0.0089	0.0014
P5	<b>0.1220<sup>^</sup></b>	0.0007	<b>0.0800<sup>^</sup></b>	0.0018	<b>0.0069<sup>^</sup></b>	<b>0.0227<sup>^</sup></b>	<b>0.0041<sup>^</sup></b>
KW $H$	44,821***	214***	8,204***	6,465***	4,822***	3,118***	1,967***

Table 4: Mean toxicity scores by protest phase and dimension. Tox.=Overall Toxicity; Sev.=Severe Toxicity; Ins.=Insult; Thr.=Threat; Idn.=Identity Attack; Obs.=Obscene; Sex.=Sexually Explicit; KW  $H$ =Kruskal-Wallis  $H$  statistic; (\*)=Threat peak at P3 (Violence); (<sup>^</sup>)=peak for other dimensions at P5 (Post-Crisis); (\*\*\*) $p<0.001$ .

Cat.	F1-P	Acc-P	n-P	F1-GP	Acc-GP	n-GP
Q3: Elev.	0.97	97.8%	2,389	Insuff.	–	231
Q4: Int.	0.75	77.0%	3,351	0.82	84.6%	214

Table 5: Best-classifier weighted F1 and accuracy per ancestor generation by conversation intensity (Q3 Elevated; Q4 Intense only; GGP insufficient in both categories,  $n<80$  post-AllKNN resampling). Cat.=Intensity Category; F1-P/GP=weighted F1 for Parent (P) and Grandparent (GP); Acc-P/GP=classification accuracy; n-P/GP=sample size post-AllKNN resampling; Insuff.= $n<80$ , insufficient minority class for reliable inference. **Dominant generation:** P in Q3 (Elevated); GP in Q4 (Intense), confirmed by four classifiers converging at 97.8% for Q3-P (Logistic Regression, Random Forest, K-Nearest Neighbors, Support Vector Classifier); Gaussian Naïve Bayes collapse (14.0%) confirms genuine signal; K-Nearest Neighbours best for Q4-GP (7.6 percentage-point F1 margin over P).

conversation: Q1 Minimal ( $\leq 0.0004$ ), Q2 Low (0.0004–0.0012), Q3 Elevated (0.0012–0.0208), and Q4 Intense ( $> 0.0208$ ; top 10% of conversations reach maximum toxicity  $\geq 0.33$ , above  $\tau=0.3$ ). Q1 and Q2 yield insufficient parent-linked chains ( $n<80$  post-AllKNN resampling), a structural property of Twitter/X’s flat conversation architecture where deep reply chains concentrate in high-intensity conversations (Saveski, Roy, and Roy 2021).

In Q3 ( $n=2,389$ ), Parent predicts child toxicity at 97% F1, confirmed by four independent classifiers converging at 97.8% accuracy; Gaussian Naïve Bayes collapse at 14.0% serves as a negative control. In Q4, Grandparent F1 (82%) exceeds Parent F1 (75%) by 7.6 percentage points, an unambiguous separation confirming GP dominance, replicating the directional finding of Falade, Yousefi, and Agarwal (2024) in Reddit communities.

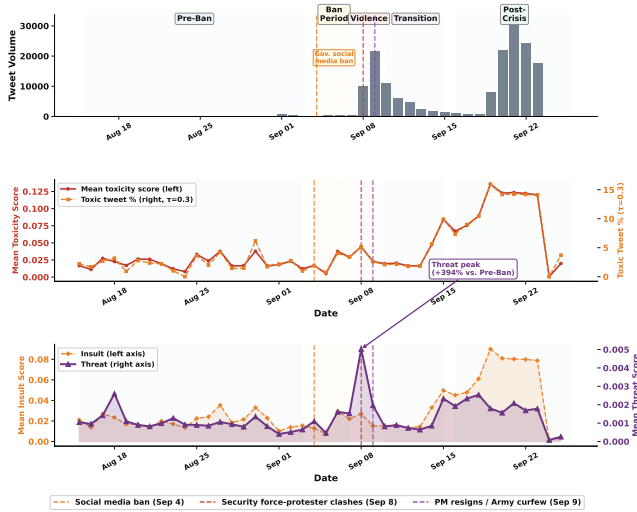


Figure 3: Temporal toxicity dynamics across Nepal protest phases (Aug.–Sep. 2025). (A) Daily tweet volume; event lines: social media ban (Sep. 4, orange dashed), security force clashes (Sep. 8, red dashed), PM resignation and army curfew (Sep. 9, purple dashed). (B) Mean toxicity score (left axis) and toxic tweet proportion at  $\tau=0.3$  (right axis). (C) Mean Insult (left) and Threat (right); Threat peaks sharply at Violence phase (+222% vs. Pre-Ban; KW  $H=6,465$ ,  $p<0.001$ ); Insult peaks in Post-Crisis, consistent with delayed blame attribution. Low P2 volume reflects the government-imposed social media ban (Sep. 4–7).

## RQ1: Cross-Lingual vs. Monolingual Toxicity Propagation

RQ1 asks: Do language boundaries amplify or attenuate toxicity propagation, and does this effect differ across toxicity dimensions? Table 6 and Table 7, and Figure 5 present the full decomposition across  $n=6,362$  monolingual and  $n=994$  cross-lingual ( $\text{switch.P}=1$ ) reply chains across five primary language transitions (English–English, Filipino–Filipino, English–Filipino, Filipino–English, Hindi–English).

Language boundaries produce a dimension-specific dual effect: de-escalation of interpersonal toxicity (five of seven dimensions dampened) while selectively amplifying physical Threat. The threat association signal ( $d = +0.145$ ), while below the conventional small-effect threshold ( $|d| \geq 0.2$ , (Cohen 1988)), is the only dimension where cross-lingual chains show elevated scores. Its significance lies not in magnitude but in direction and consistency across  $N = 172,059$  tweets: a moderation system monitoring only aggregate toxicity risks, systematically missing threatening content in cross-lingual threads. operationally critical for cyber social threat detection: a system monitoring only aggregate toxicity will *systematically underdetect* threatening content in cross-lingual threads. H1 specifically predicted Identity Attack amplification; that effect ( $d=-0.059$ ) is negligible and in the wrong direction, directly discon-

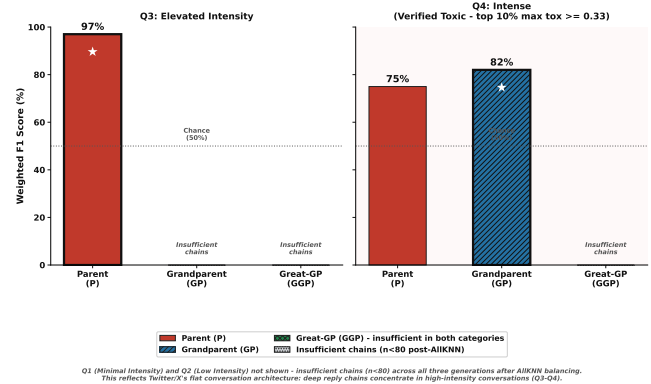


Figure 4: Best-classifier weighted F1 per ancestor generation by conversation intensity (Q3 Elevated; Q4 Intense only). Q1 and Q2 excluded:  $n<80$  post-AllKNN across all three generations, reflecting Twitter/X’s flat conversation architecture. Dashed line = 50% chance baseline; stars mark dominant generation. Q4 verified toxic: top 10% of conversations have maximum toxicity  $\geq 0.33$  (above  $\tau=0.3$ ).

Chain Type	n	Par. Tox.	Ch. Tox.	$\Delta$	% Esc.
Monolingual	6,362	0.086	0.108	+0.023	18.5%
Cross-lingual	994	0.105	0.084	-0.021	8.8%

Table 6: Chain type decomposition. Par. Tox.=mean parent toxicity; Ch. Tox.=mean child toxicity;  $\Delta$ =signed child–parent delta; % Esc.=escalation rate. Cross-lingual parents are *more* toxic (0.105 vs. 0.086) yet produce *less* toxic children (0.084 vs. 0.108), confirming genuine de-escalation rather than lower ambient toxicity. All  $p<0.001$  (Bonferroni-corrected  $\alpha=0.007$ ).

firmed H1’s primary prediction. Language boundary-crossing is associated with dimension-specific toxicity modulation rather than uniform amplification, and Threat scores show no corresponding reduction in cross-lingual chains. **H1 Verdict: Partially Disconfirmed.** Threat ( $d = +0.145$ , negligible magnitude but consistent direction across the full corpus) is the only dimension showing elevated scores in cross-lingual chains, constituting a covert threat signal masked by lower interpersonal toxicity scores (Insult:  $d = -0.335$ ; Toxicity:  $d = -0.223$ ).

## RQ2: Temporal Generational Dominance Across Protest Phases

RQ2 asks: Does the dominant generational parent shift systematically across protest phases? Table 8 and Figure 6 present phase-level results. H2 predicted Parent dominance during active protest and GGP emergence post-crisis. Results confirm H2 with unusual clarity: Parent dominates three active phases (P1, P3, P4) at  $F1=97-98\%$ , and GGP achieves the highest F1 (95%) in Post-Crisis, a clean, unambiguous temporal transition.

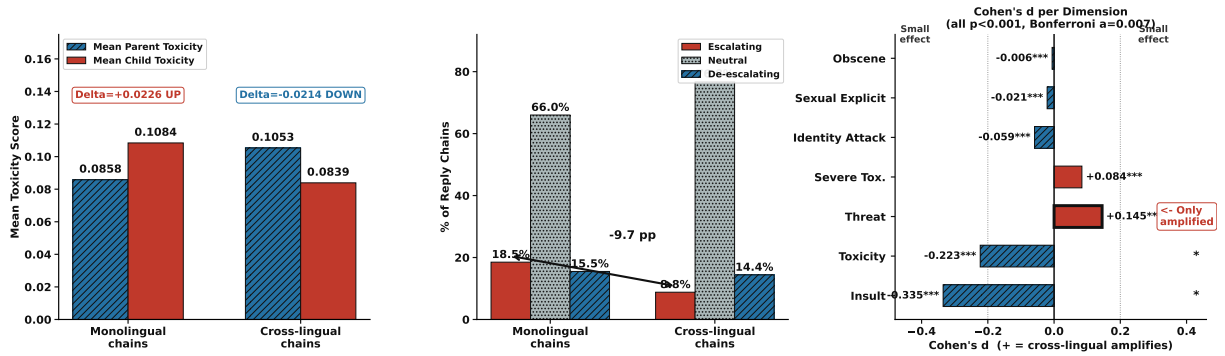


Figure 5: RQ1 cross-lingual toxicity propagation ( $n=6,362$  monolingual;  $n=994$  cross-lingual chains). (A) Mean parent and child toxicity by chain type; cross-lingual parents are more toxic (0.105) yet produce less toxic children (0.084), confirming genuine de-escalation ( $\Delta=-0.021$ ). (B) Escalation direction: monolingual 18.5% vs. cross-lingual 8.8% (gap= 9.7 percentage points). (C) Cohen’s  $d$  per dimension (Cohen 1988) (Bonferroni  $\alpha=0.007$ ; all  $p<0.001$ ); dashed lines mark small-effect threshold ( $|d|\geq 0.2$ ); Threat ( $d=+0.145$ ) is the sole amplified dimension.

Dimension	Cross	Mono	$d$	Size
Insult	0.036	0.092	-0.335***	Small
Overall Toxicity	0.084	0.108	-0.223***	Small
Threat	0.010	0.004	+0.145***	Negligible <sup>^</sup>
Severe Toxicity	0.002	0.001	+0.084***	Negligible
Identity Attack	0.003	0.005	-0.059***	Negligible
Sexually Explicit	0.004	0.005	-0.021***	Negligible
Obscene	0.029	0.030	-0.006***	Negligible

Table 7: Cohen’s  $d$  effect sizes per dimension (cross-lingual vs. monolingual child toxicity; Bonferroni  $\alpha = 0.007$ ; all  $p < 0.001$ ) (Cohen 1988). Small =  $|d| \geq 0.2$ ; Negligible =  $|d| < 0.2$ . (<sup>^</sup>) Threat ( $d = +0.145$ ) is the only dimension showing elevated scores in cross-lingual chains; its operational significance lies in direction and consistency across  $N = 172,059$  tweets, not magnitude alone. Ordered by  $|d|$  descending.

**Post-Crisis Great-Grandparent Shift.** As acute crisis resolves, the structuring influence on toxicity shifts from immediate provocation to deep conversational ancestry. GGP achieves F1= 95% vs. Grandparent (94%) and Parent (91%), a 3.3 percentage-point reversal consistent with crisis communication theory’s post-crisis sensemaking phase (Seeger, Sellnow, and Ulmer 2003), in which communities re-anchor discourse to pre-established interpretive frames rather than reacting to proximal stimuli.

**Ban Period (P2) Analysis.** P2 yields only 8 toxic chains from 161 total parent-linked chains, insufficient for reliable ML inference. Parent dominance is inferred from the convergent P1/P3/P4 pattern; excluding P2, Parent dominance ranges F1= 97–98%, confirming H2 without reliance on the Ban Period result.

**Threshold Robustness.** All primary findings hold at  $\tau=0.5$ . GGP dominance strengthens to a 4.2 percentage-point margin (GGP F1: 0.9312 vs. Parent F1: 0.8891, up from 3.3 pp at  $\tau=0.3$ ). Parent domi-

nance across active phases holds at F1= 95–97%. Cross-lingual de-escalation is confirmed (monolingual escalation 15.9% vs. cross-lingual 7.3%). Insult dampening holds ( $d=-0.291$ ); Threat amplification direction is preserved ( $d=+0.119$ ). All findings are therefore independent of the  $\tau=0.3$  calibration choice. Chain-linked subsets (Parent: 7,356; Grandparent: 650; GGP: 197) represent 4.3% of the corpus—a structural property of Twitter/X’s flat architecture, not a data limitation (Saveski, Roy, and Roy 2021). **H2 Verdict: Confirmed.** Parent dominates P1, P3, P4 (F1= 97–98%); GGP achieves F1= 95% in Post-Crisis vs. Parent 91% (3.3 pp margin, strengthening to 4.2 pp at  $\tau=0.5$ ).

## Discussion

### Cross-Lingual De-escalation and the Threat Exception (RQ1)

Cross-lingual reply chains show lower toxicity scores in association with language boundary-crossing, rather than the amplification H1 predicted. Monolingual chains escalate at 18.5% versus 8.8% in cross-lingual chains; Insult ( $d = -0.335$ ) and Overall Toxicity ( $d = -0.223$ ) both show dampening, extending Saveski, Roy, and Roy (2021)’s structural analysis by establishing language switching as a conversation-level toxicity modulator. Three non-mutually exclusive mechanisms are consistent with the evidence: *cognitive friction* (processing overhead interrupting the hostility-escalation cycle (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015)), *audience shift* (cross-language replies activate cross-community registers suppressing in-group aggression), and *self-selection* (users crossing language boundaries may be structurally less invested in in-group conflict). Distinguishing these requires future experimental work.

The Threat exception is the most practically urgent finding. Threat is the *sole* dimension showing elevated scores in cross-lingual chains ( $d = +0.145$ ,  $p < 0.001$ ),

Phase	Tweets	F1-P	n-P	F1-GP	n-GP	F1-GGP	n-GGP	Dom.
P1: Pre-Ban	3,517	0.98	261	Insuff.	63	Insuff.	–	P*
P2: Ban Period	1,808	Insuff.+	161	Insuff.	18	Insuff.	–	Inferred P
P3: Violence	31,902	0.97	780	0.95	98	Insuff.	–	P*
P4: Transition	27,823	0.98	533	Insuff.	70	Insuff.	–	P*
P5: Post-Crisis	107,009	0.91	5,621	0.94	401	<b>0.95</b>	197	<b>GGP*</b>

Table 8: Generational parent dominance by protest phase (best-classifier weighted F1, AllKNN-undersampled subsets). Dom.=dominant generation; Insuff.= $n < 80$  post-AllKNN; (+) P2: only 8 toxic chains, P dominance inferred from P1/P3/P4 convergence; (–)=no chains recovered. Threshold robustness ( $\tau = 0.5$ ): GGP dominance in P5 strengthens to 4.2 pp (GGP F1 = 0.9312 vs. P F1 = 0.8891); P dominance in P1, P3, P4 holds at F1 = 95–97%. Best classifiers: LR for P1, P3, P4, P5-GP, P5-GGP; KNN for P5-P.

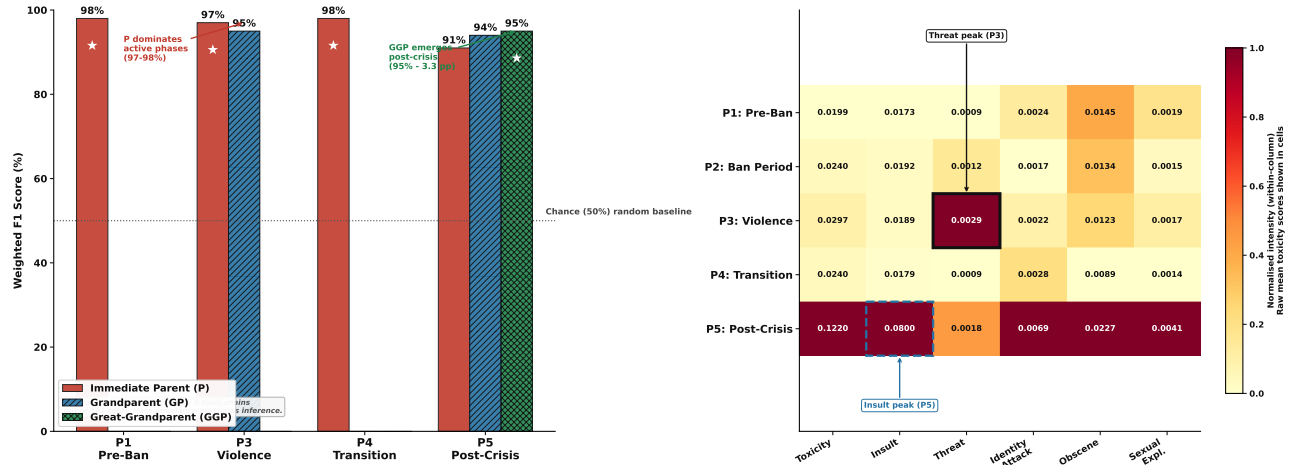


Figure 6: RQ2: Generational dominance across protest phases. (A) Best-classifier weighted F1 per generation per phase; stars mark dominant generation; P2 excluded (8 toxic chains only; Parent dominance inferred from P1/P3/P4 convergence). Parent dominates P1, P3, P4 at 97–98% F1; GGP emerges exclusively at Post-Crisis (P5: 95% vs. Parent 91%; margin = 3.3 pp). (B) Mean toxicity by phase and dimension (Table 4; colour = within-column normalised intensity); bold border: Threat peaks at Violence phase (P3); dashed border: Insult peaks at Post-Crisis (P5), consistent with delayed blame attribution.

operating independently of the general pattern – a dissociation that renders aggregate toxicity scores a *systematically misleading* signal in cross-lingual threads. Moderation systems flagging only high overall toxicity will miss the most dangerous content precisely in conditions generated by platform bans, diaspora engagement, and multilingual crisis discourse. Cross-lingual chains therefore warrant dedicated dimension-specific Threat monitoring independent of aggregate toxicity level.

### Generational Shift as a Crisis Phase Marker (RQ2)

The temporal shift from Parent (P) dominance across active phases to Great-Grandparent (GGP) dominance in Post-Crisis is the most structurally novel finding. During active phases, toxicity is provocation-driven – users react to the immediate parent. In Post-Crisis, the structuring influence shifts three turns upstream, consistent with crisis communication theory’s sensemaking phase (Seeger, Sellnow, and Ulmer 2003): communities

engage in retrospective narrative construction, retrieving earlier interpretive frames rather than reacting to proximal stimuli. The concurrent Identity Attack surge in P5 (mean = 0.0069 vs. P1 baseline = 0.0024) deepens this picture: ethnic and group-targeting rhetoric intensifies *after* violence subsides, consistent with post-crisis blame attribution dynamics. GGP dominance combined with elevated Identity Attack in P5 presents a coherent pattern: post-crisis discourse is organized around pre-established antagonistic frames directed at identifiable groups - an architectural requirement absent from current commercial moderation pipelines.

### Contributions Relative to Prior Work

Three contributions emerge. First, the association between language boundary-crossing and lower toxicity scores is established as a structural property of multilingual protest discourse, reframing multilingualism from a data quality challenge (Fortuna and Nunes 2018) to a substantive conversational variable. Sec-

ond, Falade, Yousefi, and Agarwal’s generational parent framework (Falade, Yousefi, and Agarwal 2024) is extended from Reddit to Twitter/X, revealing temporal phase-dependence of generational dominance that Reddit’s atemporal analysis could not expose. Third, a Threat-specific cross-lingual elevation mechanism is documented – absent from prior toxicity literature – with direct implications for platform moderation of cyber social threats in multilingual protest discourse.

## Limitations

Five limitations bound interpretation. **Tree depth:** only 4.3% of tweets yield a recoverable parent (7,356 chains); GGP chains represent 0.1% ( $n = 197$ ), reducing statistical power at deeper levels, though detecting effects in Twitter/X’s shallow trees constitutes a conservative test relative to deeper Reddit trees (Falade, Yousefi, and Agarwal 2024). **Model coverage:** Detoxify multilingual was not trained on Hindi or Nepali; near-zero Nepali scores reflect transfer limitations and Nepali findings are excluded from substantive interpretation. **Relevancy:** corpus mean relevancy is 28.2 (range 0–100); all tweets are retained for reply-chain completeness but findings may partly reflect off-topic discourse. **Causal inference:** all findings are observational and associational; the cross-lingual toxicity patterns cannot establish whether language boundary-crossing drives the observed score differences or whether structurally less hostile users self-select into cross-lingual exchanges. **Self-selection:** disentangling interaction effects from user disposition requires within-user analysis, comparing toxicity across monolingual and cross-lingual replies by the same user – a direction for future work.

## Conclusion

We examined cross-lingual toxicity propagation and temporal generational influence in 172,059 tweets from the 2025 Nepal GenZ protest across five crisis phases, using a conversation tree reconstruction, seven-dimension multilingual toxicity scoring, seven ML classifier families, and non-parametric testing with Bonferroni correction.

Two findings emerge. Cross-lingual chains show lower interpersonal toxicity scores while exhibiting selectively elevated Threat ( $d = +0.145$ ,  $p < 0.001$ ) – a covert threat signal masked by lower aggregate toxicity scores, partially disconfirming H1. Immediate Parent features dominate toxicity prediction across active protest phases ( $F1 = 97\text{--}98\%$ ), shifting to Grandparent dominance in Post-Crisis ( $F1 = 95\%$  vs. Parent 91%; 3.3 pp margin, strengthening to 4.2 pp at  $\tau = 0.5$ ), confirming H2 and extending crisis communication theory’s sensemaking framework to computational toxicity analysis.

Two moderation interventions follow: cross-lingual chains warrant dimension-specific Threat monitoring independent of aggregate toxicity scores; post-crisis

phases require ancestry-aware systems tracking frames established multiple turns earlier. Future work will extend this framework to multi-platform analysis across Telegram, Reddit, and synthetic temporal networks.

## Acknowledgements

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-23-1-0011, W911NF-24-1-0078, W911NF-25-1-0147), U.S. Office of Naval Research (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency, the Australian Department of Defense Strategic Policy Grants Program, Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment, and the Donaghey Foundation at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

## References

- Aleksandric, A.; Saha Roy, S.; Pankaj, H.; Wilson, G. M.; and Nilizadeh, S. 2024. Users’ Behavioral and Emotional Response to Toxicity in Twitter Conversations. In *Proc. ICWSM*, volume 18, 29–42.
- Bento, P.; Aquino, Y.; Buzelin, A.; Dutenhefner, P. R.; Chagas, A.; Estanislau, V.; Dayrell, L.; Malaquias, S.; Santana, C.; Locatelli, M. S.; Pappa, G. L.; Meira, W. J.; and Almeida, V. 2025. Mapping Hate and Extremism: How Brazilian Reddit Communities Spread Toxic Discourse. In *Proc. ICWSM Workshops (CySoc 2025)*.
- Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial Behavior in Online Discussion Communities. In *Proc. ICWSM*, 61–70.
- Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum, 2nd edition.
- Colizzi, C.; Della Sala, A. A.; Fenza, G.; and Gajewski, L. 2025. Investigating Coordinated Inauthentic Behavior on Alternative Platforms During the 2024 U.S. Election. In *Proc. ICWSM Workshops (CySoc 2025)*.
- Falade, T. C.; and Agarwal, N. 2025. Learning Hierarchical Moral Foundations for Interpretable Toxic Intent Classification via Weighted Probabilistic Soft Logic. In *Proc. IEEE Int. Conf. Tools Artif. Intell. (ICTAI)*, 1295–1299.
- Falade, T. C.; and Agarwal, N. 2026. Examining Generational Influence in Online Toxicity: Context-Dependent Patterns in Health and Political Discourse. In *Social, Cultural, and Behavioral Modeling*, 213–222.
- Falade, T. C.; Yousefi, N.; and Agarwal, N. 2024. Toxicity Prediction in Reddit. In *Proc. AMCIS 2024*.
- Falkenberg, M.; Zollo, F.; Quattrociochi, W.; Pfeffer, J.; and Baronchelli, A. 2024. Patterns of Partisan Toxicity and Engagement Reveal the Common Structure of Online Political Communication across Countries. *Nature Commun.*, 15: 9560.

Fortuna, P.; and Nunes, S. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, 51(4): 1–30.

Gao, Y.; Qin, W.; Murali, A.; Eckart, C.; Zhou, X.; Beel, J. D.; Wang, Y.-C.; and Yang, D. 2024. A Crisis of Civility? Modeling Incivility and Its Effects in Political Discourse Online. In *Proc. ICWSM*, volume 18.

Guaraldi, G.; et al. 2022. Drivers of Hate Speech in Political Conversations on Twitter. *Social Media + Society*.

Hanu, L.; and Unitary team. 2020. Detoxify. <https://github.com/unitaryai/detoxify>. Accessed: 2025-01-01.

Human Rights Watch. 2025. Nepal: Unlawful Use of Force During ‘Gen Z’ Protest. <https://www.hrw.org/news/2025/11/19/nepal-unlawful-use-force-during-gen-z-protest>. Accessed: 2025-12-01.

Myers-Scotton, C. 1993. *Social Motivations for Codeswitching: Evidence from Africa*. Clarendon Press.

Nagar, S.; Gupta, S.; et al. 2022. Capturing the Spread of Toxicity on Twitter. In *Complex Networks and Their Applications X*.

Qayyum, H.; Ilyas, I. A.; et al. 2023. A Longitudinal Study of the Top 1% Toxic Twitter Profiles. In *Proc. ACM Web Science Conf*.

Saveski, M.; Roy, B.; and Roy, D. 2021. The Structure of Toxic Conversations on Twitter. In *Proc. WWW ’21*, 1086–1097.

Seeger, M. W.; Sellnow, T. L.; and Ulmer, R. R. 2003. *Communication and Organizational Crisis*. Praeger.

Tajfel, H.; and Turner, J. C. 1979. An Integrative Theory of Intergroup Conflict. In Austin, W. G.; and Worchel, S., eds., *The Social Psychology of Intergroup Relations*, 33–47. Brooks/Cole.

Wikipedia contributors. 2025. 2025 Nepalese Gen Z Protests. [https://en.wikipedia.org/wiki/2025\\_Nepalese\\_Gen\\_Z\\_protests](https://en.wikipedia.org/wiki/2025_Nepalese_Gen_Z_protests). Accessed: 2025-09-30.

Wilson, D. L. 1972. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Trans. Syst., Man, Cybern.*, 2(3): 408–421.

## Paper Checklist

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**. We analyse publicly available Twitter/X data to advance understanding of toxic discourse propagation in crisis contexts. No individuals are profiled; all findings are reported at the aggregate level.
- (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes**. The abstract and introduction state all three contributions with the exact effect sizes and F1 values reported in the Results section.
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes**. The Methodology section justifies each design choice: the generational framework, AllKNN resampling, the  $\tau=0.3$  threshold, and the non-parametric statistical tests.

- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes**. We discuss Detoxify’s limited coverage of Hindi and Nepali, the heavy-tailed toxicity distribution, and the relevancy score distribution in the Methodology and Limitations sections.

- (e) Did you describe the limitations of your work? **Yes**. The Limitations section identifies four explicit bounds: tree reconstruction depth, model coverage for low-resource languages, relevancy distribution, and the observational (non-causal) nature of all findings.

- (f) Did you discuss any potential negative societal impacts of your work? **Yes**. The Discussion and Conclusion note that dimension-specific threat monitoring is recommended over aggregate-only monitoring, reducing the risk of systematic underdetection in cross-lingual threads that could harm vulnerable multilingual communities.

- (g) Did you discuss any potential misuse of your work? **Yes**. The cross-lingual threat amplification finding identifies a detection gap; we recommend dimension-specific rather than aggregate monitoring to support responsible moderation design.

- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes**. Usernames are used solely for deduplication and are not reported; all findings are at the aggregate level; all random seeds are fixed (`random_state=42`); code and anonymised data will be released upon acceptance.

- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**. Yes.

2. Additionally, if your study involves hypotheses testing...

- (a) Did you clearly state the assumptions underlying all theoretical results? **Yes**. H1 is grounded in social identity theory (Tajfel and Turner 1979) with code-switching theory (Myers-Scotton 1993) as a competing prediction; H2 is grounded in crisis communication theory (Seeger, Sellnow, and Ulmer 2003). All theoretical assumptions are stated in Related Work.

- (b) Have you provided justifications for all theoretical results? **Yes**. Each hypothesis is motivated by prior empirical and theoretical literature in Related Work, and each verdict is supported by effect sizes, F1 scores, and significance tests in Results.

- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **Yes**. Social identity theory and code-switching theory offer competing predictions for H1; both are discussed in Related Work and the Discussion interprets the partial disconfirmation of H1 against both theories.

- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes? **Yes**. The Discussion identifies three non-mutually-exclusive mechanisms for cross-lingual de-escalation: cognitive friction, audience shift, and self-selection, noting that distinguishing them requires future experimental work.

- (e) Did you address potential biases or limitations in your theoretical framework? **Yes**. The Limitations sec-

- tion addresses model coverage bias (Detoxify’s limited Hindi/Nepali performance), relevancy distribution, and the observational nature of all findings.
- (f) Have you related your theoretical results to the existing literature in social science? **Yes.** The Discussion systematically relates findings to Saveski, Roy, and Roy (2021), Falade, Yousefi, and Agarwal (2024), Seeger, Sellnow, and Ulmer (2003), and Falkenberg et al. (2024).
  - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research? **Yes.** The Conclusion derives two platform interventions: dimension-specific Threat monitoring for cross-lingual chains, and ancestry-aware systems for post-crisis moderation. Future work directions are also stated.
3. Additionally, if you are including theoretical proofs. . .
    - (a) Did you state the full set of assumptions of all theoretical results? **N/A.** No theoretical proofs are included.
    - (b) Did you include complete proofs of all theoretical results? **N/A.** No theoretical proofs are included.
  4. Additionally, if you ran machine learning experiments. . .
    - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results? **No.** This is an anonymous submission; code and anonymised data will be released upon acceptance.
    - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes.** The Methodology specifies the 80/20 stratified train-test split, `random_state=42`, all seven classifier hyperparameters, AllKNN resampling, and feature standardisation.
    - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **N/A.** A fixed random seed and single train-test split are used, consistent with the replication design of Falade, Yousefi, and Agarwal (2024).
    - (d) Did you include the total amount of compute and the type of resources used? **No.** Detoxify scoring used an NVIDIA GPU; ML classification used a standard CPU. Exact compute time was not recorded but all experiments are reproducible with consumer hardware.
    - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes.** Seven classifier families guard against algorithm-specific artefacts; weighted F1 accounts for class imbalance; Bonferroni correction is applied across all seven toxicity dimensions.
    - (f) Do you discuss what is the cost of misclassification and fault tolerance? **Yes.** The Discussion explicitly identifies the primary misclassification cost: aggregate-only toxicity monitoring will systematically underdetect threatening content in cross-lingual threads.
  5. Additionally, if you are using existing assets or curating/releasing new assets. . .
    - (a) If your work uses existing assets, did you cite the creators? **Yes.** Detoxify (Hanu and Unitary team 2020) and all datasets and tools are cited.
    - (b) Did you mention the license of the assets? **Yes.** Detoxify is released under the Apache 2.0 licence. Twitter/X data is used under the Twitter/X Developer Agreement for academic research.
  - (c) Did you include any new assets in the supplemental material or as a URL? **No.** Code and anonymized data will be released upon acceptance to preserve anonymity during review.
  - (d) Did you discuss whether and how consent was obtained from people whose data you are using/curating? **Yes.** All data is publicly available on Twitter/X. No individual consent is required under standard exemptions for computational analysis of public social media data.
  - (e) Did you discuss whether the data contains personally identifiable information or offensive content? **Yes.** The dataset contains usernames (used only for deduplication, not reported) and toxic content (the subject of analysis). All findings are reported at the aggregate level.
  - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? **N/A.** NA.
  - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? **N/A.** NA.
6. Additionally, if you used crowdsourcing or conducted research with human subjects. . .
    - (a) Did you include the full text of instructions given to participants and screenshots? **N/A.** No crowdsourcing or human subjects research was conducted.
    - (b) Did you describe any potential participant risks, with mentions of IRB approvals? **N/A.** NA.
    - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **N/A.** NA.
    - (d) Did you discuss how data is stored, shared, and de-identified? **N/A.** NA.