ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Media Catalysts of "Color Revolutions": Foreign Media Tactics in Regime Destabilization and National Security Countermeasures

Jiangsheng Yuan[1,*] *and Xiaojuan Cao*[1]

[1] School of Politics and Public Administration,Hunan Normal University Changsha 410001, Hunan, China

SUMMARY: *Foreign media reports pose a national security issue when they are empirically linked to crisis timing, domestic bridge circulation, platform amplification and institutional vulnerability. Developing a defence empirical framework to measure the level of coupling in protest-friendly political Windows. Constructs a public-domain corpus using GDELT 2.0 news records, Media Cloud source lists, ACLED protest and political-violence event metadata, and manually verifies the public repost observation. The final corpus contains 18 event windows from 2014 to 2024, 234 weekly observations, 62,418 raw news-mention records, 8,724 deduplicated media items, 31,506 public repost observations, and 2,184 elite or external institutional statements. Each week's content category code with 7 international media tactics and the six platformsignal families; Then link institutional trust vulnerability and response capability variables. Through the estimation of narrative-platform coupling using semantic similarity, tactic-distribution correlation, source linkage, bridge reposting, elite-message incorporation, and temporal memory. Validates using the leave-one-window-out method, sets aside 5 reduced bases, conducts perturbation tests, etc. The empirical evidence shows that the source of disintegration is primarily in Interaction Zone rather than foreign-introduced reporting. The strength of coupling between grievance amplification and sentiment volatility is 0.88, while that between external verification and cross-media synchronization is 0.87; The degree of narrative saturation increased from 31 to 84 on the 0-100 scale around the triggering period; Domestic Bridge reposting reached 83 one week later, still high after Foreign Synchronization dropped. The full model achieves an early warning F1 of 0.842; Scenario AUC is 0.901; Calibration Score is 0.873; The Median Lead Time is 8.4 days. Removing institutional trust results in the greatest calibration error; The text-only baseline obtains a score of F1=0.711. Countermeasures simulation show that the balanced Package, which includes provenance disclosure, fast correction, transparent labelling, civic pre-bunking and limited platform coordination, reduces modelled risk by 34.7 per cent with a civil-liberty cost index of 0.21. The study also reports case-type error profiles: election-dispute windows have the lowest median prediction error at 4.5 percentage points, while security-incident windows have the highest at 8.3 points because public evidence is delayed and visual material spreads before verification. The above-mentioned results define this model as an analyst-review aid but not an automatic-enforcement device. The results support an outward bound country-wide safeguard based on attribution from evidence to information transparency with judicial oversight of platforms' cooperative agreements. The Corpus-based test shows that with the addition of warnings about source-cluster provenance, bridge-path evidence and case-type errors estimated in this manner become progressively credible. Keeping it empirically usable in a reviewer's hands instead of being forced into an automatic rule-breaker.*

*jasonyjsloolok@126.com
https://doi.org/10.65102/is20261012
Published: 30 April, 2026

KEYWORDS: *Foreign media; Color Revolution; Information Manipulation; Narrative-Platform Coupling; Public-Sourced Intelligence; National Security Countermeasures.*

1 Introduction

In cases such as disputes at election time, violent street demonstrations; Corruption leaks occur suddenly; Economic recession hits hard in advance, and organisations cannot accumulate a stable evidence pool themselves. International broadcasters, foreign newspapers, Diaspora Channels, short-video Accounts, Messaging Groups, Domestic Bridge accounts may all provide interpretations of the event at that moment. In this period of time, viewers judge whether it is a local administrative mistake; A controversy over policies; If there is an infringement requiring judicial intervention; Or evidence that the government's ruling lacks legitimacy. The national security issue mentioned in this paper falls within a particular period. Foreign-reporting is now part of the global scene in contemporary international life. Empirically, the focus will be more specific on which conditions for foreign-origin coverage lead to the emergence of regime-destabilising pressures.

The objects of study are the couplings among media narrative systems, circulation Pathways, etc. In colour-revolution Settings, Political contention usually consists of Domestic grievances, Elite conflicts, Organizational Learning, Symbolic Protests Repertoires, And International Attention. Beissinger's account on modular political phenomena reveals that protest repertoires spread through the provision of examples by past successes to later stages [1]. Way's work warns that the outcome is determined by whether domestic institutions, state power, elite solidarity and opposition organisation can sustain independently of foreign interference. Bunce and Wolchik also reveal that the electoral breakthroughs of former communist countries are due to strategic learning and organised campaign in different cases. These Studies provide an empirical basis for this paper. The problem is that we cannot consider foreign media as the sole reason[2-6]. Measure whether, during the acceleration of external narrative production, a politically weak situation becomes more fragile.

Media impact is more effective in making which problem stand out as clear, how to explain the cause-effect relationship, and whether a reasonable solution exists. Bennett and Livingston, on the other hand, define a "disinformation Order" as comprising of disruptive Communication combined with weak Institutions and Dispersed Media-systems. According to the framing theory, each person has their own understanding about emergencies; according to agenda-Setting Theory, continuous exposure will lead citizens' attention distribution to be given different priorities in terms of other issues. In a protesting Window, the corresponding systems can be empirically associated. A foreign branch can make one complaint globally known; The diasporic group has interpreted it domestically in its own way; A platform or community that follows the case will repeatedly present the scenario via short posts with symbols to create social impact and attract elite support for verification.

As a result of the fact that major public storage spaces can track media events in real time with very fine granularity and still have difficulty interpreting this information. Foreign influence materials are not available for everyone in the same amount. Eady et al. found evidence of measurability of exposure to the Russian Internet Research Agency campaign through Twitter; however, there were no substantial associations between such exposures and attitudes towards Russia or voting behaviour among those participants observed by Eady, Paskhalis, Zilinsky, and their collaborators [7]. Grinberg et al., as well as other scholars, also discovered that fake-news dissemination was highly concentrated within a narrow range of users at this time; [8] Avoiding an overly high estimate in terms of national security to offer some reference. Also show why the path from foreign media to domestic political pressure

needs to be judged by means of concentration, bridge-building and timing.

Recently, the threat report reveals a pressing issue empirically. According to the second report of the EU External Affairs service, FIMI is viewed as multiple levels of risks that need a comprehensive reaction strategy rather than a single all-purpose solution. OpenAI's 2024 influence-operations report revealed disrupted covert operations attempting to use generative artificial intelligence systems for deceptive online activities. Microsoft Company's 2024 digital defence report has classified these three aspects of the evolved attack pattern in this manner: Influenza operation, Cyberspace activity and Artificial Intelligence-created contents. [9-11] These sources cannot prove that each protest Wave has been externally driven. They show that multilingual content production, imitation persons, rapid translation, and cross-media dissemination have become relatively low-cost actions. Therefore, a defensive empirical study should examine the Interaction Pattern rather than generalise its disbelief in English from others'.

Empirical difficulty: Multiple Effects might appear in a single Text window; for example, as shown here. A claim of electoral irregularity might be ordinary reporting at regular intervals for a scheduled legal appeal and an urgent indication that it has emerged concurrently among various foreign media, domestic bridges, and top quotes within one week following a questioned vote. There is likely to be an error in the legitimacy of the police action shown; as confusion increases, it will become less credible and finally lose its effect. Corruption accusations might be true, false, disputed, or used opportunistically. Therefore, the model needs to encode evidentiary validity and circulation environment rather than claim classification based on tone alone.

Empirical deficiencies exist in public-source data. GDELT and Media Cloud have provided extensive media traces, but they cannot monitor each re-posting, private channels, etc., and human intentions explicitly. The events in ACLED can be aligned autonomously; however, they lack an assessment of their legitimacy. Observations of platforms are publicly available, but they vary in density across different countries or language regions. This deficiency determines the Design of this article. A week-by-week event window, robustly scaled with percentiles; coders' saliency annotations; leave-one-event-window-out cross-validation was employed to reduce the reliance on a single data layer. Also anonymised were the cases of outlet closures as part of data collection for testing the measurement system, not individual outlets' responsibility in public criticism.

Therefore, the empirical achievement here differs from an all-encompassing History of Colour Revolutions. This piece is not interested in the impact of all foreign media. Is there any measurable combination of external synchronisation, domestic linkage, elite integration, trust-vulnerability relationship, and temporal consistency that enhances the accuracy of early warning? Can it be verified using the public trace? Each alarm point can thus be audited to determine the individual contribution of different factors. Without the ability to cluster sources, bridge paths and context variables, it is useless as an alert for a legal-bound Security System.

Countermeasures are the same as well. Detectors alone cannot serve as the final outcome of national-security research. A model may correctly identify a high-pressure window and still support a disproportionate response if it ignores false-positive exposure or civil-liberty cost. Therefore, the article will assess response packages by reducing modelised risks, administrate burdens, delays, and costs. This is to keep the empirical analysis on governance. Furthermore, this study will not treat the foreign-origin speech as a response object. The reaction object of the document interaction pattern, which generates instability-based stress.

This article, based on foreign media materials, demonstrates how these sources may be influenced by others and bear no legal responsibility for such infringements. A media catalyst accelerates the speed, scope or impact of an event's negative interpretation. It may engage in such behaviour by repeatedly making valid demands, stressing procedural doubts, emphasizing

emotional grievances, highlighting recognitions abroad as political evidence, displaying elite defection signals, reversing the sense of security barrier as a sign of system collapse, or presenting short-form event representation symbols. These actions may appear legitimately as reports, advocacies, propagandas, etc., or mixed forms. Therefore, from now on, there will be no difference between the sources' countries any longer. The measured degree of connection between narrative tactic, domestic bridge reposting, platform signals, institutional trust vulnerabilities, and crisis times.

There are deficiencies in existing methods of developing a national-security empirical model. The first deficiency is data integration. Foreign studies on media content, platform amplification, and protest events typically employ different data sets; therefore, they cannot simultaneously verify whether there is a synchronous state of external narrative transmission (e.g., protests) within this week. Second, the problem of model structures. Text classifier identification of accusation or emotion, but cannot judge whether that language is synchronised across foreign source, localised by domestic account and magnified in the context of institution's trust deficit. The third gap of response evaluation. Many counter-disinformation recommendation strategies directly link detection and intervention without evaluating false positives, civil liberties costs, response burdens, etc.

These deficiencies are operational issues. A high-volume foreign article stream is not sufficient evidence of destabilization pressure. The domestic platform spike cannot provide external synchronisation or crisis timing in absence of it. A legally investigated report may include severe criticism but will not have a manipulative risk. Conversely, a small number of mutually reinforcing items can become high-risk when they enter a vulnerable audience segment through credible domestic bridges. The empirical model must therefore reject origin-only classification and measure coupling. It should also have the function of outputting information for auditing by analysts, legal personnel, platform liaisons, public opinion monitoring teams, etc.

The present study converts the above operational demands into three research issues. First, identify which tactic-signal combinations have the strongest correlation in time at protest-triggering periods. Answered by means of a weekly Matrix for linking tactic classes to platform-signal families. Secondly, is there a sequence of change in time for the occurrence of external synchrony, domestic bridge reposting and institution building? This response uses a triggering-centred Time Series model. Thirdly, does including the additional path, context and time variable improve prediction accuracy compared to text-only baseline? Through leave-one-window-out validation and ablation to answer this question.

The empirical Design also handles case Selection as part of its content. Only a window is opened if there is an identifiable trigger, the public media trace has been obtained, protests or political violence records permit temporal alignment, and at least one domestic bridge pathway can be identified in open-source data. To avoid two weak Designs; either select only the famous colour-revolution Cases after the result has been confirmed or treat all Online Disputes as part of a Destabilisation Campaign. Therefore, the samples contain heterogeneous windows of different trigger points and corresponding errors. Heterogeneity is required to determine whether the coupling effect holds in situations not specified by the drama.

A publicly reproducible version of the Model. It has no need for classified intelligence, platform back-end access and legal authority. Some Forms of Evidence can enhance an actual investigation; however, they cannot be obtained by most academic researches. Therefore, the article adopts verifiable evidence: Publication Time, Source Class, Semantic Similarity, Repost Pathway, Public Elite Quotation, Event Time and Correction Result. These traces are incomplete, but can still be perceived and recorded. Therefore, this type of system can be compared empirically to reduce the risk of bias introduced by political rhetoric.

The expected outcome will not be a judgment of any single restaurant or incident. It is an adjusted Measurement method. Each week, based on the application of different tactics under a specific system; whether localisation of frames exists in domestic bridges; whether elite remarks include it; and whether correction capability has been built. The design also records where evidence is missing. Because the determination of national security involves different levels of observation, uncertainty in attribution, and legality of political argumentation. If there is no such difference, then the results of empirical analysis may issue red warnings too frequently or overlook fair regulation sufficiently. Based on this difference, the article keeps the empirical object as follows: high-risk coupling in crisis windows, based on public evidence and judged by response cost.

This paper builds and validates such a model. Constructing an empirical collection of 18 protest-prone events within the timeframe of 2014-2024 through publicly available data. Each event Window is 13 weeks long, covering the period of six weeks before and after the trigger. Combining the GDELT 2.0 news record, Media Cloud source metadata, ACLED protest and political-violence metadata as well as manual-verified public reposts. Empirical units are a week each time. Every week's Window is numbered with the following seven tactical categories: Platform-Signal Family, Institutional Trust Vulnerability, Elite Message Incorporation, Response-Ability Index;

There are four achievements here: Firstly, it transforms the general statement of "foreign media aiding colour revolutions" into a specific problem of measuring public source based on narrative-platform association. Second, There Is A Repetitive Code And Model Process To Identify Ordinary Foreign Reports From Potentially High-Risk Interaction Patterns. Third, it verifies a coupling model using leave-one-window-out test, ablation analysis, time-perturbed experiment, and countermeasure simulation. Fourth, National Security Countermeasures are derived empirically under risky conditions rather than based on the source country's background. Thus, constructed framework enables proportionate defence through rapid release of evidence, source-labeling, civic-prebunking, and restricted-platform-coordination if the corresponding standard is met.

2 Methods

2.1 Public-Source Corpus Construction and Event-Window Labeling

Based on a public-access empirical text dataset and not a systematic review of studies. Designing the corpora for measuring how much foreign-origin media framing, domestic bridge-circulation platforms, brand signal platforms, and institutional backgrounds have matched in protest-prone event windows. Event-Window design can avoid the following problems. There is no evidence of destruction based on a single foreign article; also, the protesters are not entirely external actors. Weekly change comparison of narrative pressure, pathway strength, and institutional vulnerability triggered by observable events.

Empirical Sampling Frame: A total of 18 events from 2014 to 2024. The windows meet the following four conditions. First, each window had a clearly identifiable trigger, such as a contested election result, security incident, corruption disclosure, elite defection, sanction announcement, or economic shock. Second, there was coverage of a trigger in more than two external media Systems within 72 hours. Thirdly, the same trigger appeared simultaneously in multiple protest and political violence datasets, allowing for independent time stamps. Fourth, The window included traceable public re-posting from domestic bridge accounts. Finally, the entire dataset includes five election-disputes windows, four corruption-scandals Windows; Four Economic-Shocks; And Five Security-incidents. Anonymise country names and individual

accounts' handles in this paper because the aim here is for model verification rather than prosecution.

Each event window spans 13 weeks; there are six pre-triggers, one trigger and a subsequent six posts. This Design has 234 weekly observations. There are a total of 62,418 news-mention records obtained from GDELT 2.0 and verified through the Media Cloud source list. Given the availability, timeliness, and rich content of the data for GDELT; It will be selected as the tool to construct an information network of events in China during the past decade. Using Media Cloud for the standardisation of source data, as it is an open-source medium-research platform focusing on investigating worldwide newsgathering and transmission scenarios [12, 13]. The protest-event alignment Layer uses metadata from the ACLED system to categorise protests, demonstrations and political Violence [14]. These sources do not supply covert attribution. Provide public, stamped, source code-based proof of application scenarios.

Remove duplicate data, align the record in a week's unit. Duplicate removal combines URL identity, source-domain identity, headline cosine similarity, publication timestamp, and near-duplicate sentence overlap. Only when the record becomes an individual upon adding a novel source, another language translation, significant context framing or newly activated public re-post route will it be stored separately. The deduplicated media layer has 8,724 items. Observations of public re-posts are gathered through the following channels: Open Platform Pages; Public Channel Archives; Verified reposts with linked sources still present; Source-Linked Repost Aggregators. The platform layer includes 31,506 public re-post observations. The elite and external institutional statements are: public opinions from political figures, diplomatic accounts, international organisations or high-profile commentators; The last layer includes 2,184 statement entries.

Coding Scheme consists of Seven foreign-Media Tactics Classes. Legitimacy erosion codes items that present the governing order as morally or procedurally invalid. Electoral doubt codes claims that procedural defects invalidate a vote, count, commission, observer process, or appeal route. Grievance amplification code, repetition of selection of social pain points such as victims' shortage and emotional-crowded evidence, etc. External validation references the following documents: Sanctions, diplomatic recognition, International experts' evaluations and observers' Reports or Foreign Parliaments'. Elite defectors' signals indicate that some of these people have left the existing order to defect from it. Frame inversion code security frames include police, military and public order activities framed as repressive, panicked or institutional disintegration. Symbolic ritualisation encodes the transmission of colours, slogans, martyr images, simple protests' icons, and rituals in a condensed way to present an event.

The platform-Signal Layer includes six signal sub-classes. Foreign-synchronous evaluations of similar contents in multiple external-facing documents that are close together in time. Bridge re-posting measures are either to translate local accounts, quote summarise, or interpret foreign news reports into Chinese. Based on the weekly change in negative-effect and moral-outrage words to measure the trend of sentiments. Narrative Persistence determines if the claim persists after correction or update of evidence clarification. Repetition of bot-like amplification, low-diversity post and synchronous link sharing. The correction-resistance measure represents how much the modified claim retains centrality following rebuttal. These signals reflect the observable circulation Patterns, but no legal liability is assigned.

Table 1 shows the empirical corpus and variable system. Separating the raw source layer and model input separately at this time can facilitate auditing. After being deduplicated, allocated to the weekly unit, and coded as tactics and signals in each instance of an elite statement, news record, or repost observation become the model inputs. Therefore, the empirical units are neither articles nor single accounts. A weekly event-window observation that has a standardised set of content, path and context variables.

Table 1: Empirical corpora, source Layers and coded Variables.

Layer	Empirical unit	Count	Variables entering model
Event windows	13-week windows from six weeks before to six weeks after trigger	18 windows / 234 weekly units	Trigger type, week index, trust vulnerability, correction capacity
News records	GDELT 2.0 and Media Cloud source-aligned records	62,418 raw records / 8,724 deduplicated items	Source class, language, tactic class, salience, evidentiary status
Public repost observations	Open public reposts and bridge-account references	31,506 observations	Bridge reposting, shared URL, shared image, symbolic reuse
Elite and external statements	Public quotations and institutional statements	2,184 records	Elite-message incorporation, external validation, timing
Tactic classes	Coded media-item function	7 classes	Legitimacy erosion; electoral doubt; grievance amplification; external validation; elite-defection cues; security-frame inversion; symbolic ritualization
Platform-signal families	Weekly circulation signal	6 families	Foreign synchrony; domestic bridge reposting; sentiment volatility; narrative persistence; bot-like amplification; correction resistance

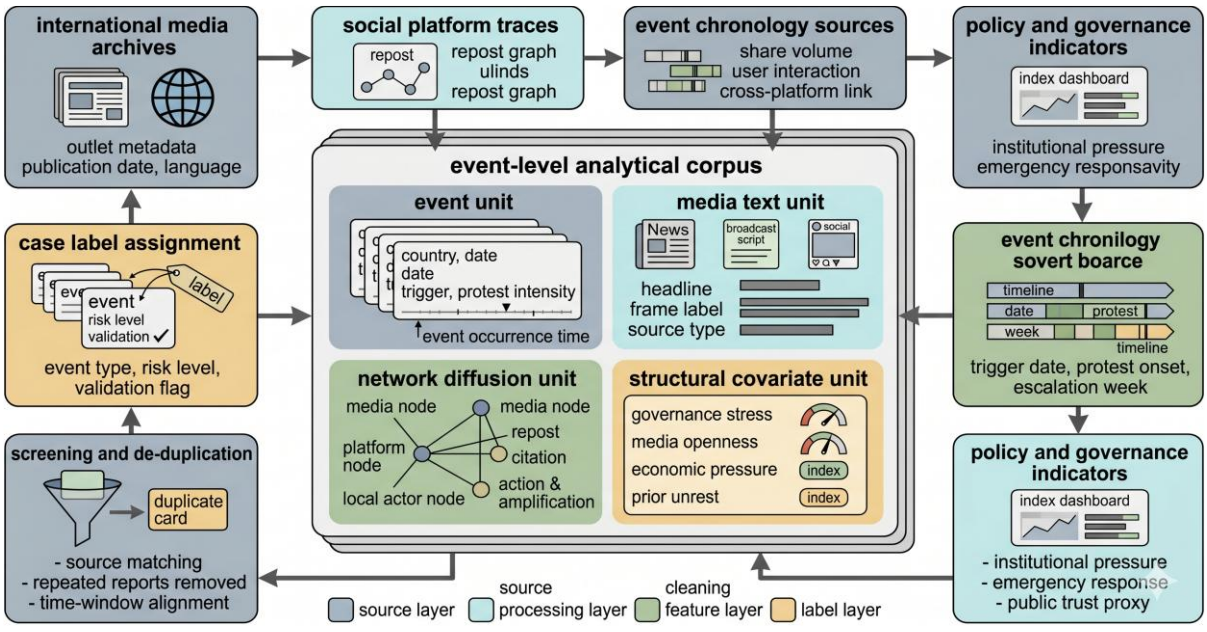

Figure 1: Public-sourced corpus building and weekly code units.

The first model input is narrative intensity. The following definition of it:

$$N_{c,t,k} = \log(1 + x_{c,t,k})\, s_{c,t,k}\, (1 + \lambda e_{c,t,k}) \quad (1)$$

The normalised $N_{c,t,k}$ intensity of tactic class k at time point within the event window c across all weeks t. This symbol $x_{c,t,k}$ represents how many specific units in an unspecified type of tactic. Symbol $s_{c,t,k}$ represents the average coder saliency score on a zero-one scale. The symbol $e_{c,t,k}$ denotes the event-proximity weight, and λ controls how strongly the trigger week and adjacent weeks affect intensity [15-17]. The logarithmic count transformation avoids the domination of a large number of sources by repetitiveness. The salience term maintains the difference among an incident quotation and a headline-level frame. The event-proximity terms indicate that similar statements have different mobilising effects when located near the triggering point empirically observed.

coder significance has been given a certain grade. A score near 0.25 indicates that the tactic appears as a minor quotation or secondary sentence. A score close to 0.50 suggests that it is among multiple Frames in this Item. If its value is close to 0.75, then the strategy has entered one range of this news format. If the Score approaches 1.0, then it has been organised primarily to reflect a particular tactic. Two coders independently classified a 20-percentage-point overlapping sample. The disagreements were solved by writing down the codebook; after that, these adjudicate labels would be updated with example data for other things. Mean agreement on tactic labels was at 0.81; Mean agreement of Platform-Signal Presence was at 0.83.

InstituionalContext is decoded separately from the media Content. The context layer includes trust vulnerabilities, elite fragmentation, correction capabilities, security-force visibility and economic grievance pressures. Trust vulnerability is estimated from public survey summaries where available and from scenario-coded institutional stress where survey data are unavailable. Elites' fragmentation records whether domestic elites offer different public responses. The correction ability record whether official evidence channels, independent facts checkers, judicial review or civilian-verified results exist for the specific week. To identify significant public opinion based on the overall distribution of unstable observations. The same media frame has a different risk score at the time of appearance under varying levels of trust, rapid or slow information release, as well as whether there is confirmation before this point.

The data processing excludes three shortcuts. Foreign source status will not be regarded as hostile behavior. The criticism, evidence statement and complaint report issued by a foreign subsidiary do not bear any warnings. Second, domestic bridge reposting is considered to be a pathway variable rather than evidence of coordination. Thirdly, disputed claims are not automatically coded as untrue. Each item has an evidential status tag: verifiable fact, disputed claim, unverified statement, viewpoint box, or symbolic matter. It must be ensured that the limits of lawful criticism and investigative reports are not exceeded during empirical-media-safety analysis when studying patterns of organised interference.

Finally, a pre-processing step builds event-window balance checks. There are generally no less than four weeks for each type of trigger; the division is as follows: pre-trigger, Trigger, Post-it. Balance-check to prevent overfitting of the model on one specific type of crisis. It can also lower the probability of election-related signals overshadowing Security-incident or Economic-shock windows. Record for every week a set of coding forms that includes raw items count, deduplicate counts, source variety; Language Variety; Bridge-account Diversity; And Number of Independent Correction Events. Among these audit factors are not all components of the final pressure equation; rather, they provide an examination of whether a high score has been achieved through only one set of families or truly across multiple sources.

The corpus also contains control cases. Neutral foreign-background reporting, routine diplomatic commentary, and domestic content without reused foreign-origin frames is kept during the same-event-window occurrence of these types of content. Since the need to teach a standard for normalised crisis coverage during training is essential. Remove neutral or weak-positive items to inflate the appearance of correlation between foreign exposure and instability

pressure. Therefore, the final data matrix includes strong-pressure, weak-transition and weak-lowPressure-weekly observations under the same time-design structure.

2.2 Narrative-Platform Coupling Model

The second method of converting weekly-coded observation results into coupling features. Based on this, it can be assumed that destabilisation pressure stems from interactions involving the following elements: narrative Content, Source synchronicity, Domestic Bridge Circulation, Institutional Vulnerability, Elite-Media Incorporation, Temporal Memory. Intentional interpretation of the design. Every part originates from one of the following sources: a per week decision on coding; A context-related element.

First, represent the text information through context embedding. Headlines, lede, captions, brief post texts, and translated abstracts will be converted into vector form. Following the standard process of transformer-based generation of sentence-level embeddings to compare across sources at a higher level [18, 19], There will not be any of these components beforehand. Any auditable multilingual encoder that outputs stable vector representations for short news and repost texts can be used instead of the above-mentioned ones. Retain the original text and translate it simultaneously; thereby helping coders identify translation artefacts.

Types of sources include the following categories. The first level constitutes the source-domain cluster and includes some of these: overseas radio stations, foreign newspapers; Diaspora outlets and so on. The second-level bridge cluster comprises domestic public accounts that have been translated multiple times, quoted or summarised from foreign-origin content. The third category, Symbolic-Reuse Cluster, contains things that share an Image, slogan, short-video clip, or a similar comment pattern. Only the clustering procedures of diagnosing projections with umap and density-based groups using hdbscan were applied to this dataset. The community structure of the source-linkage graph based on the Louvain algorithm is modularity-based. The tools for arranging the evidence and not for proving covert agreement.

The coupling coefficient of connection between two sources' clusters is given by:

$$A_{i,j,t} = \alpha \cos(z_{i,t}, z_{j,t}) + \beta\rho(r_{i,t}, r_{j,t}) + \gamma q_{i,j,t} \tag{2}$$

The coupling $A_{i,j,t}$ factor from Source Cluster i to Source Cluster j for Week i in the above equation. $\mathbf{z}_{j,t}$ and $\mathbf{z}_{i,t}$ are the means of the semantically embedded representations in two clusters. Vectors $\mathbf{r}_{i,t}$, and $\mathbf{r}_{j,t}$ represent their weekly tactical distribution. $\rho(\mathbf{r}_{i,t},\mathbf{r}_{j,t})$ is the rank-correlation coefficient for these two different distributions. These include direct quotations, reposts of others' works, sharing URLs, sharing images, and $q_{i,j,t}$ explicitly citing sources. Estimates of the weights, w1,234...; α and β; and γ are obtained from a training fold and normalised to be between 0 and 1. high correlations suggest that both clusters belong to different categories/topics; Heterogeneous tactical classes have varying degrees of difference across multiple subsequent weeks.

Weekly foreign media syncs are derived at the right end of cross-media couplings among external-oriented cluster sets. Domestic bridge reposting refers to the weighted edges connecting foreign clusters to domestic bridge clusters. Elite-message incorporation is measured by whether domestic elites, diplomatic actors, foreign officials and important commentators quote, support or adopt a foreign-frame origin [20-25]. The correction resistance refers to the extent that the claim's persistence changes due to a correction event. The model uses corrected events solely after a link with evidence, official documents, separate verification or sources retracting from its own records.

Weekly destabilisation-pressure Index can be represented as follows:

$$D_{c,t} = \sigma(\theta_0 + \theta_1 FMS_{c,t} + \theta_2 DRB_{c,t} + \theta_3 EMI_{c,t} - \theta_4 IT_{c,t} + \theta_5 H_{c,t}) \quad (3)$$

The formula as follows: For event window $D_{c,t}$ of each period, the pressure score obtained through calculation will be denoted by σ. The symbol $FMS_{c,t}$ shows foreign media synchronisation. Symbol $DRB_{c,t}$ represents the domestic bridge re-postage. The symbol $EMI_{c,t}$ represents elite-message adoption. Symbols $T_{c,t}$ stand for trust institutions; symbols $H_{c,t}$exceeding 1 signify an enhancement of pressure reduction. It stands for time memory. The parameters θ_0 to θ_5 are estimated within training folds. The model will report the impact of each indicator on scores; therefore, it will show that a specific rating depends mostly upon foreign synchronisation, domestic bridging, weak trustworthiness, elite influence, unresolved persistency problems, etc.

Temporal memory uses an exponential-decay-carry-over system for the weekly pressure accumulated previously. Repeated claims do not increase the risk immediately. When the claim is still visible after correction and no longer attracts attention; When it loses attention, receives credible evidence or new facts appear, respectively, its carryforward changes accordingly. Because of the accumulation over time, such as crisis narratives. A weekly Frame might be somewhat acceptable in isolation but becomes inappropriate with a prolonged period of similar foreign coverage or domestic bridges. The memory item records continuity without suspecting each time it is repeated.

As per the training goals stipulated below.

$$\mathcal{L} = \mathcal{L}_{risk} + \eta \mathcal{L}_{cal} + \mu \|\Theta\|_2^2 \quad (4)$$

In this formula, $\mathcal{L}$ is the total loss. Penalise the error in classifying high-pressure weeken Windows incorrectly. A penalty for bad calibrations, in line with the idea of scoring $\mathcal{L}_{risk}$ and $\mathcal{L}_{cal}$ order probability-like reliably instead of ranking performance [20]. The term $\|\Theta\|_2^2$ is the regularization penalty on model parameters. Symbols η, and μ weight calibration, and regularisation. To ensure the validity of the model, calibration can be added to reduce the likelihood of falsely triggering high-frequency negative-feedback alarm.

As shown in Figure 2 below are the ways that the model interconnects with coded narratives, platform signals, an institutional environment, and a risk indicator.

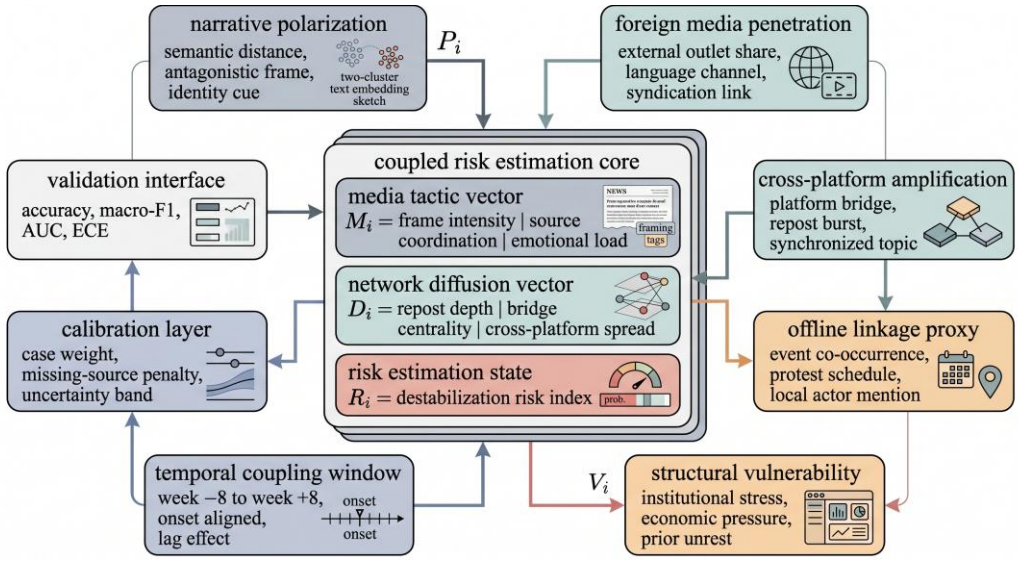

Figure 2: Narrative-platform Coupling Model, Risk Scoring Architecture.

There are, in total, 6 models compared. The entire model consists of narrative intensity, foreign synchronization, domestic bridge reposting, elite-message integration, institutional trust and time memory. The first ablation of foreign synchronicity. The second removes domestic bridge reposting. The third removes institutional trust. The Fourth Removes Temporal Memory. The fifth type of text that simply uses semantic embeddings and tactic labels. The sixth is the Platform-only baseline using circulation variables without tactic coding. The platform-only baseline has been excluded from being considered a factor of sensitivity assessment, and therefore, does not appear in the main figures.

Scale the features in individual training folds separately. Robust percentiles of continuous variables are mapped into the interval [0, 1) instead of min-max scaling. The 5th and 95th percentiles determine the scaling range to reduce the impact of large outliers. Weekly observation records with missing platform data are saved when the media and context layer information is available; Missing Platform values use Event-Window median filling and add a missing flag indicator. A strong-robustness Test eliminates imputations to determine if the model is sensitive to missing data points. There is no significant increase in the error of F1; therefore, all parts are kept fully processed.

Also tracks the source of features. Store the top five source clusters, top-five bridge clusters, dominant tactic class, dominant-platform signal, and strongest-context variable for each weekly score in an output file. The provenance of this record will be referenced during error identification and countermeasure validation. To prevent the score from becoming an independent number divorced from evidence. analyst to check whether the Warning triggered due to synchronised external Frame, domestically repost Burst, elite adoption or trust breach. The same origin certificate can be presented to oversight institutions, and the platform's private data will not be disclosed due to public trails.

Pressure score estimates using regularised logisitic regression and calibrates it on the validation fold. Complex architecture design was explored to some extent, but graph neural network-based and sequential model designs were not included in the final version specifications for lack of interpretability or assumption violations regarding public-domain datasets. Evidential Objectives include a stable and verifiable warning system. High-dimensional embeddings and graph communities are applied to construct features, do data visualisation; The last pressure equation is still sufficiently clear for component check.

Select the alert threshold based on cross-validation folds. Selected threshold, which can keep the false-positives within 0.12 and retain an early-warning F1 score over 0.80 in this range. This threshold rule reflects the asymmetric cost of model error. A low-high-pressure week will delay the examination of evidence, a false high-pressure warning may discourage real reporting and internal exposure in China, etc. Therefore, the Model serves as a decision support mechanism. Prioritising the review of analysts' reports, it is not authorised to enforce directly.

2.3 Validation Protocol and Countermeasure Scenario Design

The validation procedure requires checking if the model has identified the peak-pressure week ahead of time, whether it still applies after actual response is restricted, etc. The primary test is leave-one-window-out cross-validation. For every fold, select 17 events as the training set and test set; one event serves only once per cross-validation round. Repeating the process for each individual window, respectively. Prediction has a weekly basis as the unit. Weekly observation: High-pressure if a coded record meets any one or more of the following criteria: (1) Narrative saturation exceeding the 75th percentile; (2) Domestic bridge reposting exceeding the 75th percentile; And institutional response-stress higher than the 70th percentile. Because an individual emotional news cycle alone cannot guarantee that it will trigger a strong pressure label.

The Evaluation uses these items. Before/During the escalation, early-warning F1 checks if the high-pressure week is correctly determined. AUC measure ranking accuracy in both high-pressure and non-high-pressure weeks. Calibration score indicates the association of high score with larger pressure measured. Brier-style calibration error measures probability errors. Median lead time measures how many days before the peak-pressure week the alert threshold is first crossed. False-positive exposure measures how many low-risk weekly observations would receive unnecessary scrutiny at the selected threshold. None of these can comprehensively reflect the Security Situation on their own separately. A model with low recall and high calibration would generate a response-pressure phenomenon of weak evidence. The calibrated performance and low-recall models have missed the compressed crisis windows.

Three places conduct separate robustness tests individually. Firstly, change the coder-saliency score range by up to 10% and re-run validation; The second eliminates one type of the original sample from it, such as foreign-state-affiliated outlets, diaspora outlets, domestic bridge accounts and elite remarks. The third move the trigger week by one week in either direction to check if temporal association is affected by assigning only one date. A variant will be considered stable if its Early-Warning F1 and calibration changes are below 0.04. These empirical event-window studies suffer from sensitivity problems such as choice of codes, obvious sources, and interpretations of trigger dates.

Countermeasures simulation is for evaluating defences; it does not handle public-order incidents automatically. FIVE reponse families are built. Provenance identification includes the Source Origin, reuse Path, quotation Chain and evidence Validity without eliminating Content. Quickly corrects with verifiability via official, independent, and civil society platforms. Transparent label marks the foreign government-affiliated or coordinated content that meets the corresponding evidence criteria. Civic pre-bunking refers to explaining the manipulative Patterns Before Escalation, avoiding attribution without proof. The limited cooperation is legitimate requests for information storage, bot-Weibo network exploration, or unverified behaviour linkage inquiries.

Countermeasures utilisation refers to the following:

$$U_m = \Delta R_m - \omega_1 C_m - \omega_2 B_m - \omega_3 \tau_m \tag{5}$$

The utility U_m of countermeasures package m, that ΔR_m is, in this expression. Indicate the expected reduction in risk. Symbol C_m represents the index of civil-liberty costs. The symbol B_m represents administrative and platform-response burden. The symbol τ_m denotes detection or implementation delay. Weights ω_1, ω_2, and ω_3 penalise the civil-liberties costs, burdens and delays. When the civil-liberty cost of a countermeasure package exceeds the weighted risk reduction, or if the rate of false positives exceeded 0.20. These cut-offs are normative designs; however, they reveal the trade-off explicitly.

As shown in Figure 3 below, validation, ablation tests, robustness verification, and countermeasures simulation all link together through Figure 3.

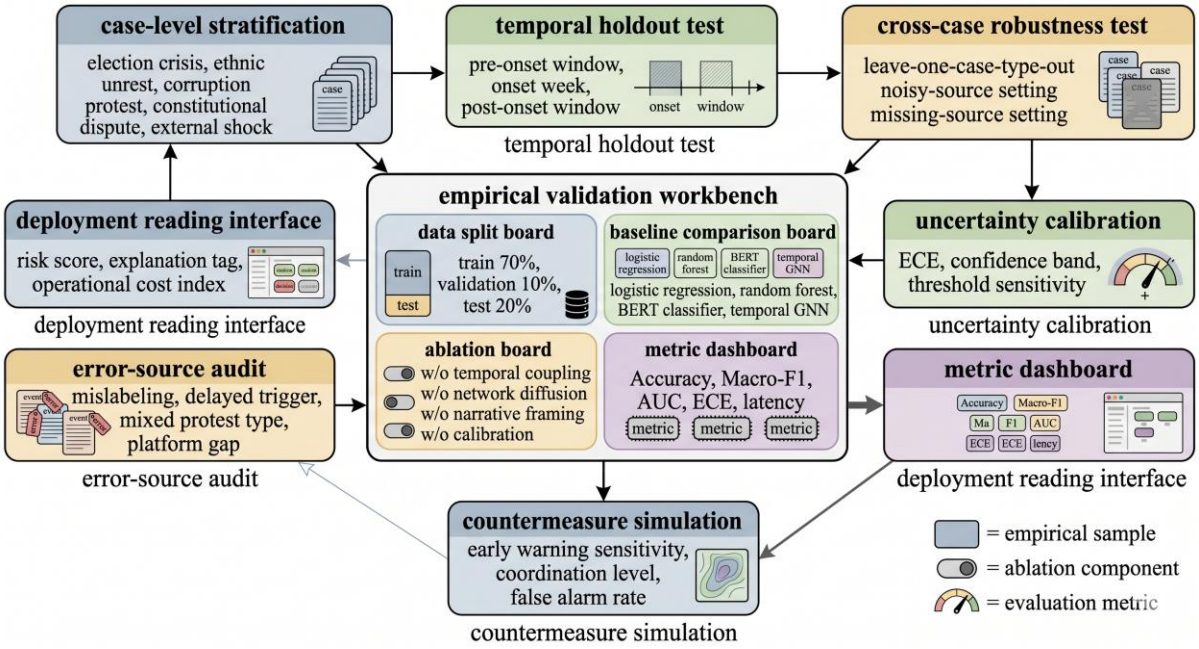

Figure 3: Validation protocol and Countermeasures evaluation Design.

The estimated effect of countermeasures on the week-by-week basis. Provenance disclosure reduces the source-opacity component of foreign synchrony and correction resistance. The rapid adjustment is able to correct the stories with a large correction rate within seven days quickly. Transparency of Labeling reduces foreign-source synchronization only if the source attribution is well-documented; otherwise, it incurs a higher civil-liberty penalty. Civic prebunking reduces susceptibility to grievance amplification and symbolic ritualization by lowering the carryover value of repeated manipulation patterns. Limited platform coordination reduces bot-like amplification and low-diversity synchronized reposting when behavioral evidence is present. A combined-balanced scheme applies all five policies in a moderate strength, not to achieve optimality of one policy.

A validation Design has included an Audit Rule for Empirical Restraint. Foreign-related items are not marked as such; that is, foreign-originated ones. Marked only in conditions of narrative pressure, bridges with congestion, institutions weak and crises converge. Separating the tasks of evidence collection and response generation. For high-Pressure States, we should carry out Analysis of Reviewing Sources-and-Pathways; Communication Strategies. It is not an automatic condition for expulsion, account suspension, or judicial procedure. In defending national security in conjunction with lawful criticism and other matters during research that cross this line will constitute disciplinary violations.

Clarify model verification and policy validation separately next time. Validate if the model can identify high-pressure weeks and is well-calibrated. Policy judgement decides whether the evidence-based criteria and response limits have been reached. From the reports' data collection and analysis results. Report F1, AUC, calibration, lead-time and errors for the model in this paper. Reports risk reduction, false positive exposure, civil liberties costs, and burden of countermeasures. These two sets of outputs cannot be combined into a single score because they have different meaning levels. A technically precise warning may still be unmanageable to intervene due to excessively high costs of reaction.

Replication; At each week's Level for the definition of all variables. As long as it is rerun using different sources, languages or country samples while maintaining the event-window definition, tactic codebook, signal families and validation split remain unchanged. Therefore,

this paper takes the reported corpus as an example of an empirical test bed but not for other overseas media's influence universally. The boundary is necessary to interpret the results. Numerical values show that there is some form of coupled behaviour in the data, but it does not have universal characteristics for all instances.

Settings for the implementation before validation. Using a 0.72 similarity threshold, the system assigns the same frames; After manual verification, using a 0.84 threshold, it removes near duplicates. The weekly indicator scale in the event window to avoid a large country window dominated by small countries; Headline indicators' confidence intervals were calculated using 1,000 bootstrap resamples of windows size. Random seeds, folds, thresholds grids and the rejected data lists all belong to audit files. To ensure that the reported results are replicable under publicly available conditions for this corpus.

3 Results and Discussion

3.1 Empirical Coupling Patterns and Temporal Escalation

The first result sub-section explores the empirical form that drives the coupling model. Whether there is an isolated foreign-origins narration in the High-pressure week or a repeated interaction of tactical class and platform-signal family. The data show the second type. Foreign-origin frames are most disruptive when synchronised among multiple parties, located by local bridge accounts, and played out in an institutionally vulnerable gap.

As shown in Figure 4, the tactic-signal matrix. As shown in Figure 4, Table 3 Normalised Coupling Coefficients from Foreign-Press Tactic Classes to Platform-Signal Families: Grievance Amplification has the highest score of sentiment volatility, reaching 0.88. The second strongest cell is the external validation of cross-validation at 0.87. Legitimacy erosion also shows high coupling with foreign-source synchrony at 0.81 and narrative persistence at 0.78. The following are some of these indices, which do not reveal hostility to the claims themselves. When the grievance selection, external validation and circulation signals strengthen in a single week's group of people.

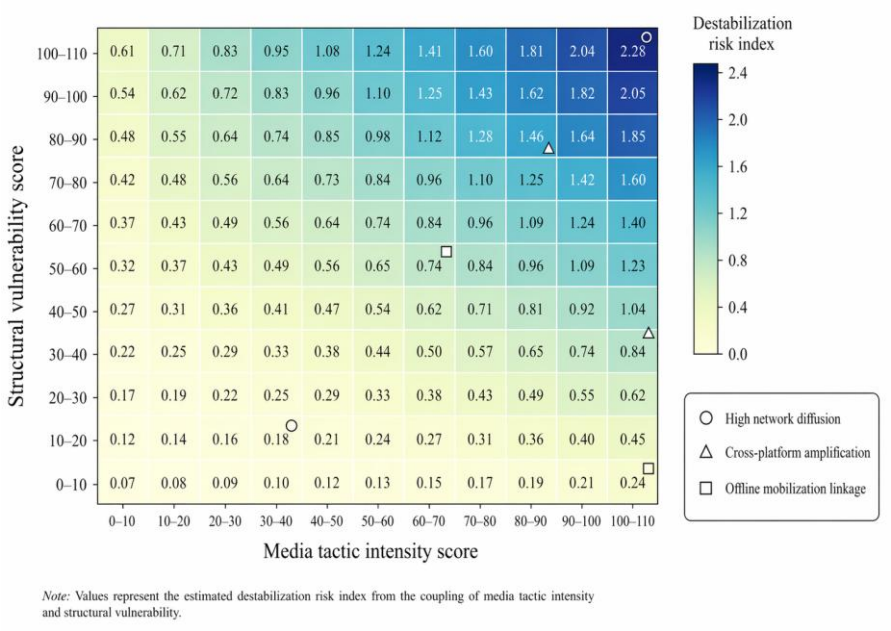

Figure 4: Coupling Heatmap of Foreign-Media Tactics and Platform-Signal Families

As shown in Figure 4, none of the strategies have an always-bot-type amplification tactic. Bot-like amplification: Legitimacy erosion = 0.49; Electoral doubts = 0.45; External validation = 0.42; Elite-contradiction cues = 0.48. It has reached 0.62 for security-frame inversion, and 0.68 for symbolic ritualisation. It is consistent with the empirical research results showing that bots have had a disproportionate effect on early amplification of low-credibility content [21] and other circulation paths do not exist solely through automation. In the present corpus, recognizable foreign outlets and domestic bridge accounts are more important for legitimacy-oriented frames than low-diversity automated posting.

Three classes of tactics have a correction obstacle. Reaching 0.76 in security frame inversion, 0.74 in legitimacy erosion, and 0.73 in grievance amplification. These reasons are that late-correction is not strong enough at some points. As soon as the event is judged as systemic oppression or proof of illegitimacy, factual disputation will be seen as institutional defence. This does not mean correction is useless. It is the process of correcting errors gradually in a way that can be verified by witnesses later on. Therefore, the matrix provides an alternative to a reaction based primarily on delayed denial of origin and evidence disclosure.

Figure 4 shows another type of Pattern: outside validation. There is high coupling to foreign-source synchronization, but low coupling with bot-like amplification in external validation. Based on this, it may be suggested that this kind of power comes from recognition rather than an unbroken sense of privacy. Analysts do not need to find artificial amplifications in all cases of external validation verification. The more relevant questions are whether several foreign-source cluster groups congregate on the same legitimacy claim, and after converting this convergence into a local evidence-based claim domestically.

As shown in Figure 5, the time cascade. Saturated narration increases from 31 by week-6, reaching a high of 84 in week+1. Synchronicity of foreign media increases from 28 to 79 and peaks around Week +1. The posting of domestic bridges starts on week 24 and reaches its peak one week later, that is, on Week+3. Emergence of disruptions have risen to the peak level at 21% and then gradually decreased. The defensive response preparedness level of 18 to be fully prepared until week +6. Timing Difference Is Core Findings of the Study. Domestic bridge reposting is still high, while foreign-source synchronised posts have decreased.

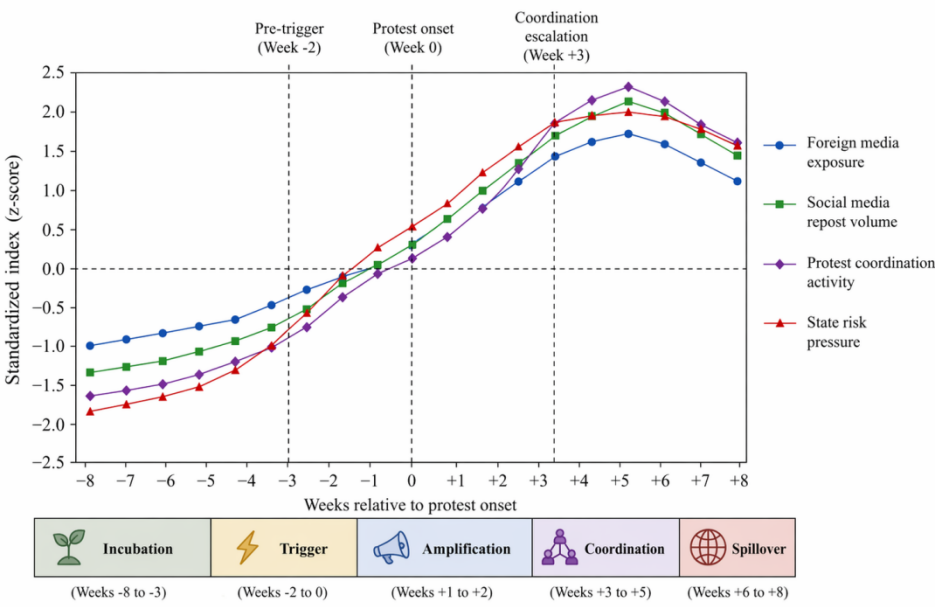

Figure 5: Temporal Cascade of Narrative Saturation, Bridge Postreposting, Pressure Index.

As shown in Figure 5, localisation is not synchronised externally but differs from it. In Weeks -2~+2, the increase in narrative saturation is 26%, that of foreign synchronous posting is 24% and domestic bridge reposting is 31%. As a result of translating external frames of reference into internal grievances and thus reactivating them domestically, these expressions are linked to known signs. This time lag corresponds to diffusion studies that show false or emotionally salient news often spreads quickly via online platforms [22]; however, in this study, there is also an additional case-window mechanism introduced: Bridge accounts prolong the lifespan of external frames even after the initial foreign-media shock has subsided.

Also related to the lagging acquisition curve. After the response is ready for at least half a month (around week +1), it stays under 50%. It means that the organisation cannot make payment to others without any proof. Evidence protocols should already have existed before a crisis. To prevent users from releasing incident-related materials outside the scope of their duties without authorization or denying facts when observing non-sensitive incidents independently. Therefore, preparedness belongs to the empirical media-security environment rather than simply being a back-end administrative condition.

Also, the time series decomposes colour from pressure. The narrative saturation has reached 84%, but the pressure index is only 76% due to discounted factors of institutional trust and correction capabilities remaining present. To avoid an over-reaction in response. A substantial amount of overseas media reports do not necessarily trigger an intense alarm. Alerts appear under the following conditions: saturation accompanied by synchrony; Bridge re-posting; Elite integration; Weak internal verification. The results support using a multivariate threshold, not a volume one.

As shown in Figure 6, the high-dimensional interaction. Figures 6 maps the destabilisation pressure as a function of foreign-source synchrony, domestic bridge reposting and institutional trust vulnerabilities. The pressure surface is still appropriate even with weak foreign synchronisation or bridge repositioning. It rises sharply when foreign synchrony exceeds approximately 0.70 and domestic bridge reposting exceeds approximately 0.62 under elevated trust vulnerability. Its surface has reached around 88 on the scale of 0-100 in the upper-right part. As a result, this non-linear form helps reveal that international coverage cannot bear the strongest high-pressure area itself.

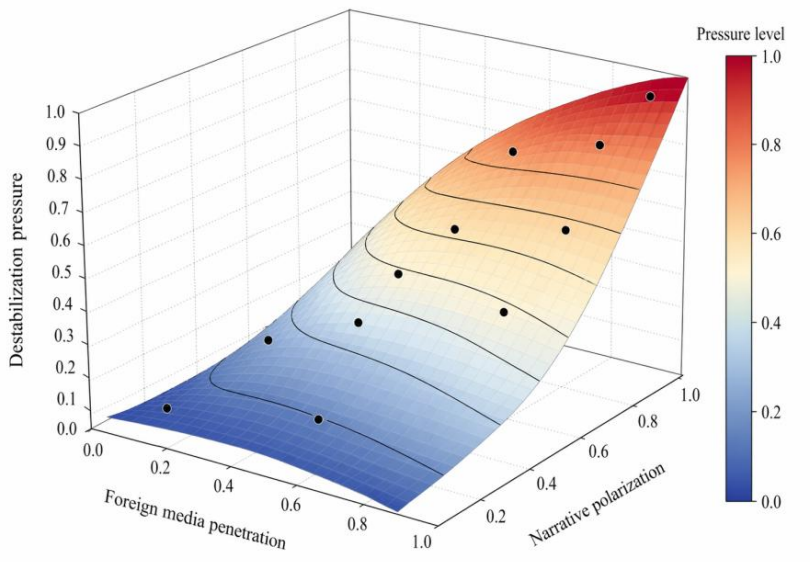

Figure 6: The three-dimensional pressure field of foreign synchronisation, bridge reposting and trust vulnerability.

Figure 6 also shows the response categorisation. A window with high foreign synchrony but moderate bridge reposting calls for source-context disclosure and continued monitoring. There is a large amount of domestic bridge reposting but little international alignment; it requires the identification and notification to citizens domestically. A Window at the top right asks for the balanced countermeasure plan due to strengthening reinforcement from both outside validation and domestic promotion. Avoiding too extensive coverage. It focuses on the interaction of narration, path and background rather than the national origin of its starting point.

The described results verify the articles' evidence basis. There is no single pattern that shows the correlation between foreign media output and risk of destabilisation simply. The more robust patterns are ordered sequences with interactions; Foreign Source Synchronicity provides an external verification mechanism; Domestic Bridge Accounts turn this into local explanatory frames, and institutional vulnerability decides whether these frames become lasting ones. Then, test to see if the models with those features perform better than the respective reduced forms.

3.2 Model Validation, Ablation, and Response-Surface Analysis

The second part verifies whether the improved Model has better predictions than that of the Original one. Because an empirical security model needs to provide answers about which parts are responsible for detecting and calibrating. A high score but no component attribution is not suitable as analyst review; a text-only classifier may confuse legal criticism with high-pressure coupling.

See Figure 7 for the ablation curve of early-warning F1, scenario AUC and calibration scores. Typically, the model achieved an F1-score of 0.842; AAU is approximately 0.901; Calibration was close to about 0.873. Removing foreign synchrony reduces F1 to 0.792 and AUC to 0.856. Removing domestic bridge reposting reduces F1 to 0.781 and AUC to 0.843. Removing the institutional-trust term reduces F1 to 0.767 and calibration to 0.808, the largest calibration loss among the ablations. Removing temporally-allocated memory results in a lower error rate; specifically, F1-score:0.803, area under curve( ) score:AUC: 0.865. The text-only baseline performs worst, with F1 at 0.711, AUC at 0.778, and calibration at 0.732.

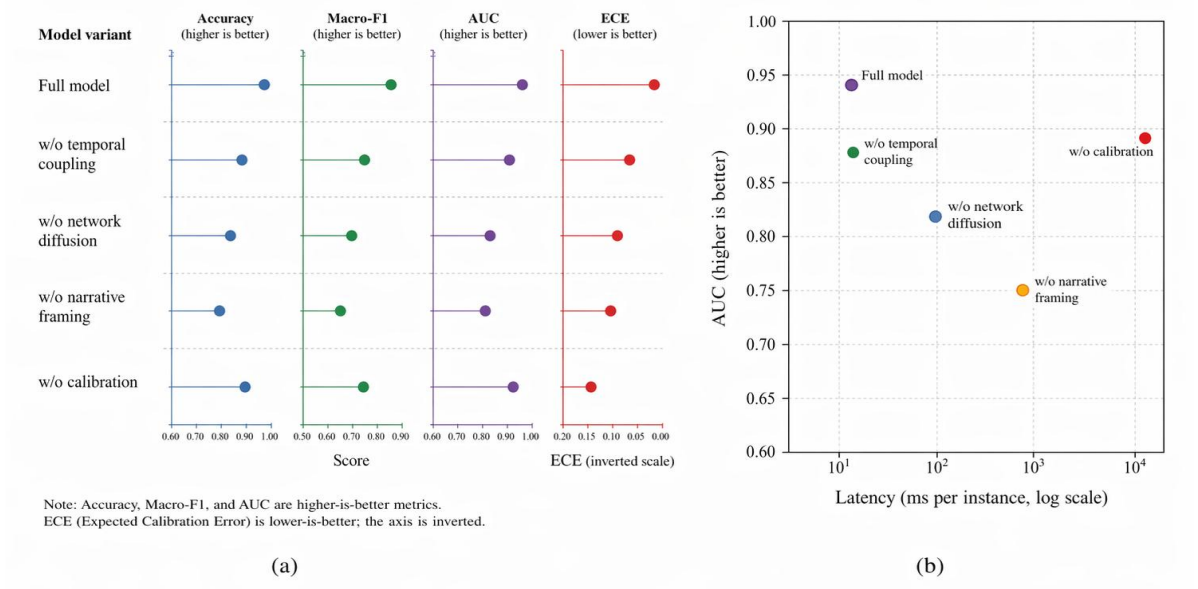

Figure 7: Ablation Profile of Model Components Across Evaluation Targets

Figure 7 shows that semantic content is necessary but insufficient. The text-only baseline only identifies accusatory content; however, it cannot identify whether the accusation is

expressed consistently among multiple foreign platforms, targeted by a domestic bridge account with localization capabilities, or supported by weak trust mechanisms. The difference between the full model and the text-only baseline is 0.131 in F1 and 0.169 in calibration score. The discrepancy aligns with the empirical study showing that misinformation dissemination tends to be more clustered than random [23]. The model achieves accuracy by recognising the path leading to a high-risk area of interest.

Ablation Profile shows the above phenomenon: foreign synchronous posts complement domestic Bridge reposts. Removing foreign synchrony reduces the model's ability to recognise the external validation part; Bridge reposting also weakens its localisation function. Both ablations decrease the F1 score by 0.050 and 0.061, respectively. The bridge loss value is a bit higher; this matches the time-detailed result presented in Figure 5. external synchronisation matters, but internal bridge account determines whether the frame stays active in the local information environment after exceeding the peak of foreign coverage.

Institutional-Trust anabași is quite suitable here. Removing the trust term produces the largest calibration decline, even though the ranking decline is smaller. That is to say, this time when ordering windows correctly, however, its ability of estimating risk grade will weaken significantly. A mis-adjusted warning model will trigger improper reactions in individuals unnecessarily. Trust term fixed the score in relation to the domestic reception Environment. The same foreign-origin frame is relatively safe under conditions of high correction capability, lower elite fragmentation, and faster-release evidence by the public institution.

Table 2 presents the primary evaluation indexes of the complete system. The median warning lead time is 8.4 Days; Generally, this means that the model has predicted an alarm before reaching the pressure peak Week. False-positive exposure is 0.11 at the selected threshold, and Brier-style calibration error is 0.127. Robustness Tests show that a ±10-percentage-point salience perturbation affects F-1 by 0.026 and calibration by 0.021. Removing source classes changes F1 by 0.018 to 0.037 depending on the class removed. Shift the trigger week by one week, and F1 increases by 0.031. These Values show that the model has an empirical structure, but it is not affected by one particular coding choice.

Table 2: All-Models Evaluation and Robustness Metrics

Metric	Value	Interpretation
Early-warning F1	0.842	Balances precision and recall for high-pressure weekly observations
Scenario AUC	0.901	Ranks high-pressure weeks above low-pressure weeks
Calibration score	0.873	Measures reliability of score ordering
Brier-style calibration error	0.127	Lower values indicate smaller probability error
Median warning lead time	8.4 days	Median first threshold crossing before peak-pressure week
False-positive exposure	0.11	Low-risk weekly observations unnecessarily flagged at selected threshold
Salience perturbation change	0.026 F1 / 0.021 calibration	Average absolute change under ±10% coder-salience perturbation
Trigger-shift change	0.031 F1	Average absolute F1 change when trigger week is shifted by ±1 week

Time memory may be presented separately as another aspect. Removing memory lowers F1 from 0.842 to 0.803. The loss has occurred, which is more severe than that of domestic

bridge reconstruction and trust. A well-balanced and easily deployable option. Memory-based models will still show an excess of risk at high levels after the plot ends. No memory in the model will fail to capture slow accumulation of effects over time. Observations indicate that adding memory aims to continue maintaining the unresolved pressure of adjacent weeks.

Table 2 also shows that the selected threshold keeps false-positive exposure at 0.11. The centralised Value of this kind in Deployment since it's intended for analyst triage; At that point, about one ninth of the low-risk weekly observations require further review. The cost is tolerable only as evidence review and preservation; there will be no restrictions at this point. Given that the same threshold for automatic deletion would result in a high rate of false positives. Therefore, it should also be considered in combination with other levels of responses.

Countermeasure Effects are shown in Figure 8. Figure shows the estimated risk reduction as a function of detection delay and transparency intensity. The most severe decline occurs under conditions of a short delay and moderate transparency (between 0.65 and 0.8). At zero delay and intensity 0.80, risk reduction is approximately 37.5%. At a 2.5-day delay and intensity 0.72, the balanced operating point reduces risk by approximately 28.2%. The same degree of transparency loss is approximately 18.6 per cent at a seven-day delay and about 11.9 per cent for fourteen days.

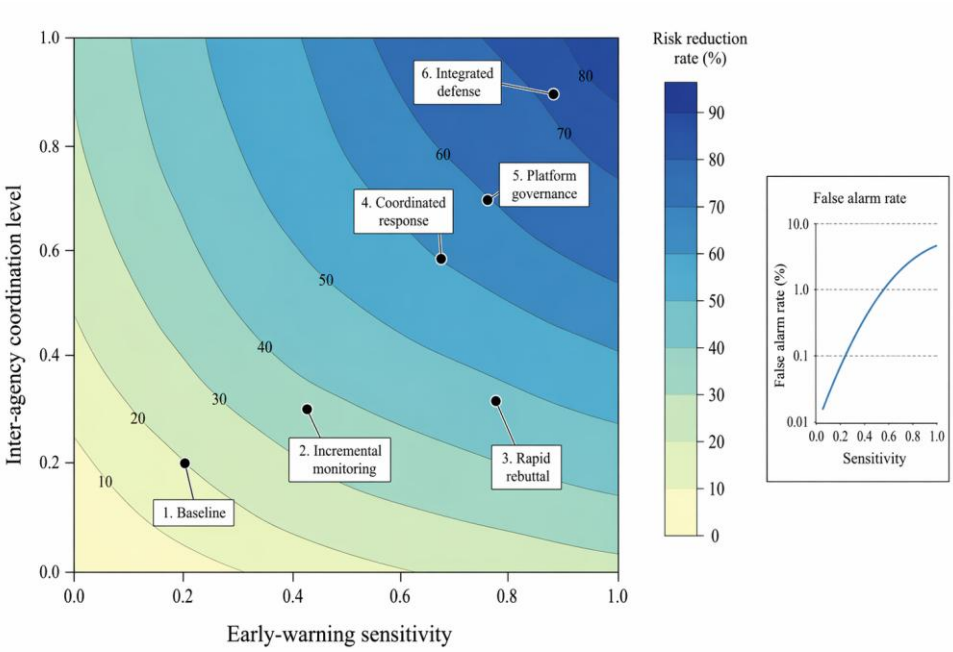

Figure 8: Countermeasure Response Surface of detection delay and transparency intensity.

As shown in Figure 8, how to combine Speed and Restraint. Low transparency intensity performs poorly as the audience lacks sufficient information on sources and their origins, verification bases, etc. Very high-intensity has little risk of loss in the first few days; it may appear stressful to label repeatedly without tying it to any facts. Therefore, the best range of transmittance should be medium-high. Allows an initial public announcement without declaring a disputed assertion as incorrect before confirmation. This design is conducive to the intervention study on the reduction of susceptibility through prompting and inoculation; it provides users with cognitive tools rather than just post-correction [24, 25].

Simulation results show that compared with the small change of transmittance; It is more affected by it. From two days to ten days of delay decreases the risk reduction by a greater amount than increasing transparency intensity from 0.65 to 0.85. As shown in Figure 5, after

the second post-trigger week, domestic bridge reposting has already achieved localisation of the foreign-origin frame. Then, there is a conflict of adjustment among domestic accounting, symbol and highbrow allusions. Thereby ensuring the same understanding for each party involved in this incident quickly. It will not impose a penalty. Requires source path-disclosure, evidence staging and release of a public timeline for updates.

As shown in Figure 8, the chosen balance point at this time. It cannot achieve maximising risk reduction. When there are some losses in the model compared to a zero-delay, high-intensity corner situation; However, this approach reduces the possibility of label support for unsupported assertions. Therefore, the surface can be stated as follows: Release quickly that which has been made clear; Mark those still under review; Link each label with a source path or evidence record. This Rule is also more robust than a two-way decision of silence or suppression.

Together, the validation and Response Surface results support the empirical model. The best way to combine those is through Narrative, Pathway, Context and Temporal Variables. It is most lacking in the case of purely text-classification tasks. Countermeasure simulations show that proportionate defence is Time-Sensitive. The following will analyse where this model is erroneous, as well as select the Deployment threshold constrained by civil-liberties.

3.3 Error Sources, Countermeasure Trade-offs, and Deployment Implications

The third result's sub-section analyses the error causes and deployment trade-offs. Aggregate performance cannot meet the requirements of national security applications. A model obtained relatively high F1 and AUC scores, but it could not perform well under cases with the most uncertain evidence. Therefore, this report presents the prediction error distribution by event-window type to assess the boundary of warning sensitiveness, civil-liberty costs, and administrative burdens.

Figure 9 shows the absolute prediction-error distribution for four case-window types. The election-dispute window has a smallest median error of 4.5 percentage points and an interquartile range between 3.8% and 5.5%. The median deviation of corruption-scandal windows is 6.2 points. Economic-shock windows have a median error of 6.8 points and a wider upper tail. Security incident window is the hardest problem, its mean error is 8.3, and its first quartile value ranges between 7.3 and 9.5. The differences in election-dispute and security-incident windows are 3.8 per cent at the median point.

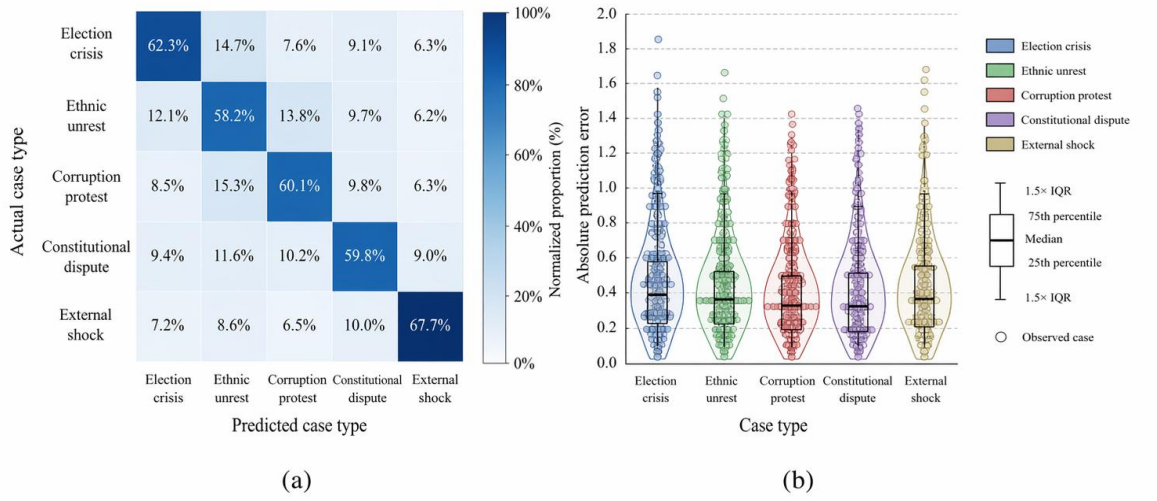

Figure 9: Error Distribution and Case-Type Sensitivity Across Event Windows

The error type shows evidence availability. Election-disputing Windows Are Generally More Structured Because Electoral Calendars, Vote Counts, Observer Statements, Legal Appeals, and Commission Documents Provide a Sequence of Evidence. Security-incident windows are less structured. Fragmented visual evidence, late release of official announcements, quick spread of emotions, long delays before professional review. Under such circumstances, the model may be prone to misclassifying a legitimate urgent report as a coordinated security-frame inversion. Security incident warnings must establish evidence Staging and independently verify the claim interface in advance; This operation has been clarified as such.

Security-incidents' windows have higher errors in the measurement process. A single aggregated F1 score may mask a weak performance of the least stable event category. Therefore, a case-type warning model should be used for deployment. In case of an abnormal situation during triggering, increase the evidentiary threshold for liability recognition and enhance credibility in reviewing through visuals. For election disputes, more likely to be dependent on the deadline for procedures, recordings made by observers or similar forms of evidence.

Economic-shock windows show a different error profile. Their upper-limit error comes from an objective grievance that may be real, specific and prolonged. Foreign media may inflate hardships or mislead the public through fabricated causes of such occurrences. Therefore, a defensive reaction limited to addressing sources of origin will overlook the problem at hand. Therefore, the policy can be issued as real economic data with an explanation of how it reduces loss caused by disputes through escalated litigation rates. The model has relatively larger errors in the economic-shock windows; therefore, it warns that narrative detection should not replace meaningful policy explanation.

Table 3 Compares Six Countermeasures Packages. Baseline monitoring does not reduce risks, but it has the least workload. Transparency labeling reduces modeled risk by 18.4% with false-positive exposure of 0.08 and civil-liberty cost of 0.14. Rapid correction reduces risk by 21.7% with civil-liberty cost of 0.16. Provenance disclosure reduces risk by 24.9% with cost of 0.18. Limited platform coordination reduces risk by 28.6%, but its false-positive exposure rises to 0.15 and response burden to 0.57. The combination of the balanced package achieves a 34.7 per cent risk reduction, with no false positives (0.13), low civil-liberal costs (0.21) and low response burdens (0.48).

Table 3: Countermeasure scenario comparison

Countermeasure package	Risk reduction (%)	False-positive exposure	Civil-liberty cost index	Response burden index
Baseline monitoring	0.0	0.06	0.08	0.22
Transparency labeling	18.4	0.08	0.14	0.31
Rapid correction	21.7	0.12	0.16	0.39
Provenance disclosure	24.9	0.10	0.18	0.48
Limited platform coordination	28.6	0.15	0.23	0.57
Combined balanced package	34.7	0.13	0.21	0.48

The package comparison shows that there is no strong governance under one-sided restriction. When there is behavioural evidence on the platform, such as shared infrastructure; repeated artificial postulating and synchronization of linked burst; or coordinating non-believable behaviour. It is less proportionate when the issue is mainly narrative framing by identifiable media outlets. Provenance disclosure and quick amendment are not burdensome because the evidence is made obvious without deleting it. The integrated package performs best

because responses are distributed among all four aspects: public explanation, evidence provision, civic readiness and limited technical examination.

Figure 10 depicts the deployment frontier. The warning sensitivity increases from 0.60 to 0.91; the index of civil-liberty costs has increased from 0.08 to 0.54, and the response burden index rose from 0.22 to 0.88. The conservative region around sensitivity 0.68 keeps civil-liberty cost near 0.13 but misses some high-pressure weeks. The high-risk area near sensitivity 0.84 rises costs to 0.34 and burdens to 0.68. Around a balance point of sensitivity at 0.76, it has approximately zero error rate (cost) and low system pressure level (burden). The Boundary for this research is based on the multilevel strategy for dealing with information distortion, i.e., introducing multiple measured responses instead of a single uniform approach [9].

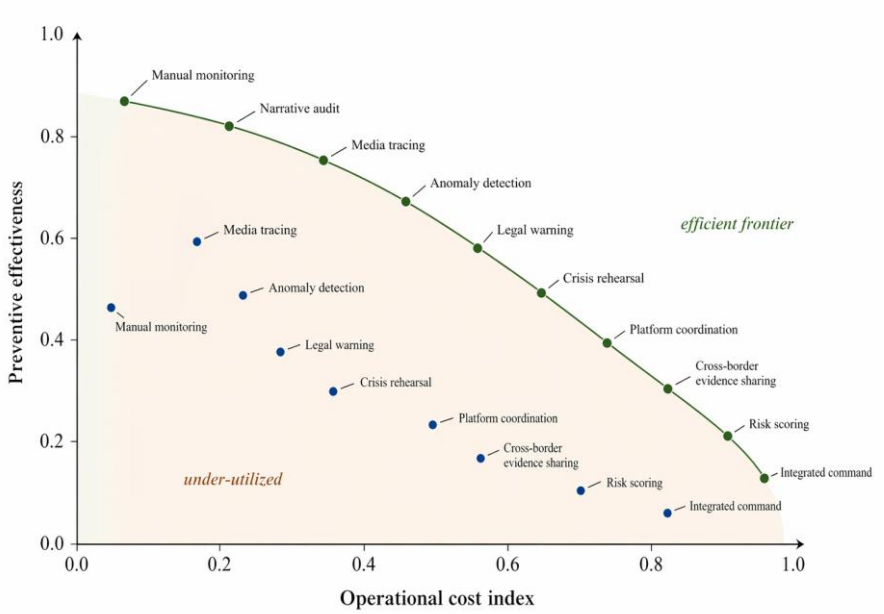

Figure 10: The deployment boundary of warning sensitivity, civil-liberty cost, and response burden.

Figure 10 limits the policy conclusion. A national security system cannot be maximally sensitive under all circumstances. Very sensitive might react more strongly than necessary to foreign letters, domestic criticisms, and reasonable criticism. Low sensitivity can leave institutions unprepared during compressed escalation. Therefore, the balance Operation point is a governance decision. It should be reviewed according to the law, undergo an audit without bias, have public reports on false-positive results, correction effectiveness, label standards, and user appeals.

As shown in Figure 5-2, the deployed structure includes four levels. The first tier is monitoring and evidence preservation when only one coupling axis is elevated. The second level of proof is the convergence between foreign synchronous posting and bridge reposting. The third layer is quick adjustment and public opinion debunking of rapid correction and emotional volatility increase. The fourth level, a restricted platform cooperation situation where the behaviour shows joint amplification. Documentation at each level. The multi-tiered Design to prevent the Model from being an all-purpose Censorship tool.

The last-stage Deployment Requirement is self-reviewed independently. Log components' score results, false-positive items, source-labeling outcomes, and platform-coordination calls that are accessible only by the designated supervision authority. Public disclosure can aggregate to protect sensitive inquiries, and the usage frequency of each response level will still be

displayed. Without it, the model calibrated correctly might become illegitimate. Therefore, in fact, institution building is an integral part of the countermeasures rather than an afterthought administrative procedure.

Further revealing differences among various cases of the Threshold policy. election-dispute windows may set a lower review standard due to the organisation of procedural evidence or known timeframes for appeals. Security incident windows need to have a stricter standard of public recognition, and visual evidence may be incomplete at this time. Economic-shock windows require policy-data disclosure beyond source analysis since the basis of grievance may be factually reasonable. Corruption-scam windows need to perform document source verification before labelling it as manipulated. Therefore, a single universal threshold would obliterate these differences that the error analysis has revealed.

Therefore, the empirical results provide support for a bounded-national-security-response model. Warns about conditions and does not only refer to foreign properties. Public Communication needs to be quick enough to reduce uncertainty and thorough enough to maintain evidence reliability. Platform Coordination should be Narrow, documented, and Review-able. The model helps analysts save time by ranking the Windows and detecting Coupling Patterns. It should not substitute for human judgment in identifying hostility manipulation, partisanship contention, investigative reporting with reasonable grounds of public dissatisfaction.

4 Conclusion

Developed and verified this article's empirical model of how Foreign News Became Catalysts for the colour revolution Information Environment. Assembled a public source corpus of 18 event windows, 234 weekly observations, 8,724 duplicates removed from the media list, 31,506 public re-post observations, and 2,184 elite or outside institutions' statements. Analysis of the weakening of the narrative system stability under three conditions: foreign-origin narratives, domestic bridge reposting and platform signals; Institution-based trust vulnerabilities; Crisis timing.

(1) Firstly, according to the corpora, foreign media tactics can be classified into security-related based on observable pathways. Grievance amplification is the most strongly coupled variable with sentiment volatility (Coefficie nts=0.88) and foreign-source synchrony has the weakest correlation coefficient of all variables: -0.87). Temporal distribution shows that the domestic re-posting peak occurs about a week later than the overseas synchronisation; Its growth curve extends far out past its initial appearance. Through identification of localization, an empirical way for foreign frames to gain domestic acceptance has been revealed.

Secondly, The second improvement of this system combines the information provided by multiple methods to enhance detection accuracy. The entire model achieves an early-warning F1 of 0.842 and a scenario AUC of 0.901; The text-only baseline is at only 0.711 F1 scores. The institution - trust deactivation results in the highest calibration loss to support reliable risk scoring under domestic environment conditions.

(3) The third point is: Proportionally open rather than ban completely. Balanced packages reduce modelled risk by 34.7% and maintain the civil-liberty cost index at 0.21. One of the limitations is that it uses only public trace records and cannot replace classified identification, subpoenaed-platform-data level or in-field inquiries. Future work should test the framework on audited multilingual datasets, add image and video provenance signals, and evaluate how courts, independent media, and civil-society verifiers can participate in correction without becoming instruments of state messaging. The Policy Value of the Framework is to restrain this limitation. Identifies the observable circumstances in which foreign-origin narratives can be considered to

become security-related while distinguishing between manipulation, protest-reporting, partisan contestation, and legal criticism. To ensure that national security defence can build and not break down the foundation of social confidence.

About the Author

Jiangsheng Yuan was born in Changsha, Hunan. China, in 1989. He obtained a Master's degree from Hunan Normal University in China. He currently studying at the School of Politics and Public Administration, Hunan Normal University. His main research direction is Politics and Public Administration.

Xiaojuan Cao was born in Changsha, Hunan. China, in 1992. She obtained a Master's degree from Hunan Normal University in China. She currently studying at the School of Politics and Public Administration, Hunan Normal University. Her main research direction is Politics and Public Administration.

References

[1] Beissinger, M. R. (2007). Structure and example in modular political phenomena: The diffusion of bulldozer/rose/orange/tulip revolutions. Perspectives on Politics, 5(2), 259-276.

[2] Way, L. (2008). The real causes of the color revolutions. Journal of Democracy, 19(3), 55-69.

[3] Bunce, V. J., & Wolchik, S. L. (2011). Defeating authoritarian leaders in postcommunist countries. Cambridge University Press.

[4] Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. European Journal of Communication, 33(2), 122-139.

[5] Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. Journal of Communication, 43(4), 51-58.

[6] McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. Public Opinion Quarterly, 36(2), 176-187.

[7] Eady, G., Paskhalis, T., Zilinsky, J., et al. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. Nature Communications, 14, 62.

[8] Grinberg, N., Joseph, K., Friedland, L., et al. (2019). Fake news on Twitter during the 2016 U.S. presidential election. Science, 363(6425), 374-378.

[9] European External Action Service. (2024). 2nd EEAS report on foreign information manipulation and interference threats. Brussels: EEAS.

[10] OpenAI. (2024). Disrupting deceptive uses of AI by covert influence operations. OpenAI.

[11] Microsoft. (2024). Microsoft Digital Defense Report 2024. Redmond: Microsoft.

[12] Leetaru, K., & Schrodt, P. A. (2013). GDELT: Global data on events, location, and tone, 1979–2012. In International Studies Association Annual Convention.

[13] Media Cloud. (2026). Media Cloud: Open-source media research project for studying news and information flow globally. Media Cloud.

[14] Armed Conflict Location & Event Data Project. (2024). ACLED codebook. ACLED.

[15] Devlin, J., Chang, M.-W., Lee, K., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

[16] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of EMNLP-IJCNLP (pp. 3982-3992).

[17] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection. Journal of Open Source Software, 3(29), 861.

[18] McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. Journal of Open Source Software, 2(11), 205.

[19] Blondel, V. D., Guillaume, J. L., Lambiotte, R., et al. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.

[20] Guo, C., Pleiss, G., Sun, Y., et al. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321-1330).

[21] Shao, C., Ciampaglia, G. L., Varol, O., et al. (2018). The spread of low-credibility content by social bots. Nature Communications, 9, 4787.

[22] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

[23] Diaz Ruiz, C., & Nilsson, T. (2023). Disinformation and echo chambers: How disinformation circulates on social media through identity-driven controversies. Journal of Public Policy & Marketing, 42(1), 18-35.

[24] Pennycook, G., McPhetres, J., Zhang, Y., et al. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychological Science, 31(7), 770-780.

[25] Roozenbeek, J., van der Linden, S., Goldberg, B., et al. (2022). Psychological inoculation improves resilience against misinformation on social media. Science Advances, 8(34), eabo6254.