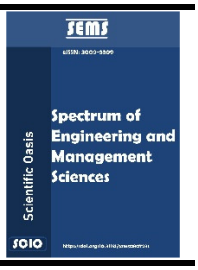
SCIENTIFIC OASIS

Spectrum of Engineering and Management Sciences

Journal homepage: www.sems-journal.org
ISSN: 3009-3309



Platform Governance as Decision Support: A Governance Risk Index for Twitter/X During the 2024 U.S. Presidential Campaign

Abdullah Önden[1,*]

[1] Department of Computer Engineering, Istanbul University, Faculty of Computer and Information Technologies, Istanbul, Türkiye

ARTICLE INFO

Article history:
Received 10 January 2026
Received in revised form 22 March 2026
Accepted 23 March 2026
Available online 24 March 2026

Keywords:
Platform Governance; Algorithmic Accountability; Democratic Discourse; Social Media Governance; Governance Risk; Digital Platforms; Toxicity Amplification; Platform Accountability

ABSTRACT

Online platforms have emerged as essential infrastructure for democratic deliberation, crisis communication, and public health messaging. Nonetheless, governance of these spaces remains largely retrospective and principle-based, without concrete evaluative criteria. This article operationalizes six key principle-based elements of platform governance into concrete clause components with quantifiable triggers, to support proactive application. We develop a composite Governance Risk Index (GRI), integrating empirically defined decision thresholds for four risk components: dispersion, drift, inequality, and toxicity. We estimate toxicity levels from a stratified subsample of the dataset labeled using the Perspective API, aggregated to the daily level. Our results show that toxicity is positively correlated with engagement ($r = 0.52$, $p < .001$), as replicated in our data ($r = 0.49$, $p < .001$), in line with algorithmic amplification dynamics that platform governance must account for. We further find that decision thresholds differ meaningfully between clause components and risk components. A contextual benchmark against 2015–2016 multi-platform baselines reveals that the 2024 Twitter/X environment exhibits substantially different statistical properties in engagement dispersion and toxicity prevalence, highlighting the need for context-dependent calibration of platform governance thresholds. Over a 30-day out-of-sample holdout period, validated exclusively against independently verified external events, the GRI classified governance-event days under retrospective validation with 85.1% accuracy. Under a supplementary labeling scheme that incorporates platform-internal anomaly criteria, accuracy reaches 90.0%, representing an improvement of about 13 percentage points over single-metric baselines, highlighting the advantages of multivariate, multi-faceted operationalization in proactive platform governance.

1. Introduction

Social media are now central to democratic deliberation, crisis communication, and public health messaging [1]. As platforms assume infrastructural status, their governance has become ever more important. Platforms must strive for optimization, but do no harm, balance business and social goals,

* *Corresponding author.*
E-mail address: abdullah.onden@istanbul.edu.tr

https://doi.org/10.31181/sems41202669

and human and algorithmic functions [2]. The governance challenges arising in these socially vital digital spaces demand analytic tools that translate the language of normative claims into the language of actionability. Social Media Analytics (SMA), the use of computation to extract insights from user-generated text, operates at the governance boundary. SMA is a vehicle for platform optimization, but also entails risks including privacy harms, exacerbating inequalities, and algorithmic bias [3,4].

Three main shortcomings exist with respect to current platform governance strategies. Firstly, they are largely reactive: platforms respond to threshold violations post hoc rather than proactively avoiding threshold violations [5]. Secondly, existing governance frameworks offer normative guidelines, including transparency, accountability, and fairness, but lack mechanisms for converting these abstract notions into actionable rules [6,7]. Finally, the governance literature typically employs data from a prior period of platform evolution, namely pre-algorithmic-feed and pre-pandemic, so that the findings are of limited applicability to current issues of toxicity amplification, disinformation, and fake news, and reduced research access due to API restrictions [8].

The motivation for this paper is to bridge the gap between normative governance principles and operational decision-making in platform governance. While the literature has extensively discussed what platform governance should achieve, there is a lack of concrete, quantitative tools that practitioners can use to detect governance-relevant events in real time and to trigger appropriate responses. This study is motivated by the need for an empirically validated, composite governance index that translates abstract governance principles into actionable decision boundaries. Accordingly, we pose the following research questions:

i. Can a composite governance index, combining engagement, sentiment, inequality, and toxicity risk components, outperform single-metric approaches in identifying governance-relevant events on social media platforms?
ii. How do empirically derived governance thresholds differ across risk dimensions and platform contexts, and what does this imply for threshold calibration?
iii. How can normative platform governance principles be operationalized into concrete, implementable decision rules with quantifiable triggers?

The three-fold objective of this research is to:

i. define a composite Governance Risk Index (GRI) that combines the different risk signals into a unified governance signal;
ii. test the GRI index using a large-scale, real-world dataset of 22M Twitter/X posts related to the 2024 U.S. Presidential Campaign;
iii. translate normative governance principles into concrete governance clauses with quantifiable decision boundaries and escalation triggers.

1.1 *Contemporary Governance Challenge*

Three developments have changed the platform governance landscape since 2020. First, algorithmic amplification prioritizes content that optimizes engagement, which has been linked to toxicity and polarization [9]. Second, the ecosystem characteristics that underlie governance frameworks vary across platforms and over time; as a point of reference, we include a benchmark of 2024 Twitter/X data against our 2015 to 2016 multi-platform data (Facebook, LinkedIn, Google+), and note descriptive discrepancies in engagement inequality, toxicity prevalence, and sentiment valence. We stress that these discrepancies represent the joint influence of platform, domain, and

time, and thus cannot be interpreted as temporal trends (see the Methodological Caveat). Third, the 2018 and 2019 "APIcalypse" [5] has made empirical platform governance research increasingly difficult at a moment when platforms are growing more opaque, and therefore calls for governance frameworks that can be validated with available data.

1.2 *Research Gap and Contribution*

Although the governance literature provides normative goals [6,7] and design models [10], two gaps remain:

i. *Gap 1 – Lack of quantitative decision making tools for governance*. Existing models specify who (authorities), what (data lifecycle), and why (accountability, fairness) of platform governance, but not how governance decisions are actually made. This is not a tools gap, but a criteria gap: the literature describes the logic of governance, but not the practice of governance. Although composite indices have been developed in other contexts, including the Digital Economy and Society Index (DESI) for national digital readiness and algorithmic auditing frameworks [11], no index has developed quantitative, real-time decision boundaries for applying platform governance principles to social media analytics data.
ii. *Gap 2 – Limited empirical validation*. Many governance proposals illustrate their logic using synthetic data or last year's data. Few proposals are empirically validated: tested on real-world platform crises to compare expected and actual outcomes. Research on the effectiveness of content moderation [12] has shown the value of empirically validating governance proposals, but this has not been extended to composite governance indices.

We fill these gaps with three contributions:

i. *Contribution 1: Operationalized Governance Mechanisms*. We operationalize platform governance principles as six empirically-calibrated governance clauses that include decision thresholds, decision rights, and escalation procedures to offer guidance for practitioners. This translation of governance principles into concrete and executable decision rules is the primary contribution of this study to the practice of platform governance.
ii. *Contribution 2: Governance Risk Index—Composite Governance Index*. We develop and validate a four-component composite index that combines engagement, sentiment, inequality, and toxicity into a single governance signal, and evaluate its event-detection performance on a holdout period.
iii. *Contribution 3: Contemporary Empirical Validation*. We validate the index on 22 million Twitter/X posts related to the May–July 2024 U.S. Presidential Election [13], evaluating its performance on real-world events including presidential debates, campaign events, and crisis communication incidents. Importantly, while this study uses publicly available datasets, the scientific contribution lies not in data collection but in the novel analytical framework: the construction of a composite governance index, the derivation of empirically calibrated governance thresholds, and the operationalization of governance principles into actionable decision rules. The use of established, publicly available datasets ensures reproducibility and enables independent verification of our results, which we consider a methodological strength rather than a limitation.

1.3 *Practical Significance*

For platform managers, the GRI offers a proactive risk assessment tool. We apply empirically-derived thresholds to categorize governance risk, a crisis detector to differentiate between noise and signal, and an explainable governance tool to provide an auditable rationale for governance intervention through the index components. For regulators, we provide a tool to translate policy objectives (e.g., purpose limitation under the GDPR, algorithmic explainability) into quantifiable metrics. For researchers, we demonstrate a scalable, post-hoc empirical approach to validating governance outcomes using publicly available data to overcome access barriers to proprietary platforms.

1.4 *Theoretical Foundations*

Our work integrates three streams of literature. Information systems governance [10,14] offers broadly applicable, stage-dependent control frameworks that remain to be operationalized in the context of platform analytics. Digital responsibility [6,7] offers principles of stakeholder accountability that remain to be operationalized in computational terms. Platform governance [2] identifies algorithmic accountability needs that remain to be formalized in mathematical terms. Our work contributes to all three streams by developing an empirically validated framework.

1.5 *Paper Organization*

The paper is organized as follows. Section 2 reviews the literature on platform governance, social media analytics, digital footprints, accountability, inequality, and contemporary platform dynamics, and develops the study hypotheses. Section 3 describes the methodology, including research design, data sources, variable measurement, construction and validation of the GRI, and analysis framework. Section 4 reports the empirical results, including contemporary Twitter/X dynamics, GRI classification performance, contextual benchmarking, and governance clause operationalization. In Section 5, the theoretical and practical contributions of the study are discussed. In Section 6, the limitations of this study are outlined. Finally, Section 7 presents the conclusion.

2. Literature Review

2.1 *Social Media Analytics: From Insight Generation to Governance Object*

Originally, SMA was introduced as a capability to generate insights from UGC for customer engagement, sentiment analysis, influence identification, and anomaly detection [15]. Early SMA research focused on business value creation through customer insights [15], brand reputation [16], competitive intelligence [16], and so forth.

The status of SMA changed from business value to governance object for three reasons. First, the digital traces of social interactions constitute a valuable resource for behavioral inference and attention management [17] that is produced by, but whose value is appropriated and utilized through, aggregation and inference by a different stakeholder, leading to information asymmetries that are hard to resolve [18]. Second, analytically-driven personalization can create business value and exacerbate social inequality at the same time: the optimization of engagement may inherently disadvantage users with sparse digital footprints [4]. Third, analytics can contribute to, or help to resolve, platform failures (e.g., disinformation outbreaks, harassment, and election interference) depending on design choices [19].

2.2 *Digital Footprints: Assets, Liabilities, and Governance Imperatives*

Digital footprints are information that is left behind by users when they create content, interact with others, use a device, and are tracked as they move across platforms [3]. Digital footprints are a

key resource for platforms in that they allow platforms to provide a better user experience. Simultaneously, they are a threat to users since they reduce the users' level of privacy, expose them to profiling, and are difficult to contest [23]. We know from empirical research that users differ in their awareness of digital footprints and control over information disclosure [17,24], complicating the process of informed consent. Three issues emerge in the governance of digital footprints:

i. *Purpose Drift* – Footprints collected for one purpose (e.g., optimizing engagement) are used for another (e.g., creditworthiness, job screening) without the knowledge and consent of the user [25].
ii. *Ecosystem Diffusion* – As footprints flow across vendors, cloud services, and data brokers, it becomes difficult to assign responsibility to a specific actor [26].
iii. *Algorithmic Inference* – Inferred footprints (i.e., information about users that is inferred by machine learning algorithms but not actually provided by them) pose unique challenges for the governance of digital footprints [27].

The GRI framework proposed in this paper addresses these challenges by calling for the development of purpose registries (that list all the uses to which footprints are being put), vendor assurance clauses (that place limits on the extent to which third-party analytics providers can access footprints), and distributional monitoring (that monitor how different groups of users are being impacted by footprint-based decision-making).

2.3 *Platform Accountability and Algorithmic Governance*

Three failures of governance have been identified in the literature on platform accountability:

i. *Transparency Failures* – The lack of transparency of algorithms, as well as the proprietary status of data, makes it difficult for external actors to monitor platforms [2].
ii. *Amplification Failures* – Algorithms designed to maximize engagement result in the disproportionate promotion of disinformation and hate speech [9].
iii. *Accountability Failures* – The distributed character of platform ownership, algorithm design, and content moderation makes it difficult to hold actors accountable [1].

There are two streams of literature that are of direct relevance to our discussion of the operationalization of platform governance. The first is the literature on content moderation that has shown that platforms deploy a combination of automated and human moderation techniques that are shaped by complex policy frameworks [12]. The second is the literature on algorithmic auditing that has developed techniques for the external monitoring of platforms. However, most of this literature has focused on the problem of discrimination and bias rather than on the broader set of governance risks faced by platforms [11]. Recently, there has been an emerging body of research that has sought to quantify the problem of amplification. For instance, in a study of Twitter, Wu *et al.* [9] found that toxic content receives substantially more engagement than neutral content ($r = 0.52$, $p < .001$). For governance purposes, this implies that platforms need to override engagement-based metrics with metrics that are aimed at minimizing harm. As we will discuss later, our proposed GRI incorporates a measure of engagement that weights the toxicity of the content. Taken together, these three failures of governance, transparency failures, amplification failures, and accountability failures, highlight the need for operational instruments of governance that can complement normative instruments that do not have clear decision-rules.

2.4 *Inequality, Inclusion, and Distributional Governance*

A growing body of literature has shown that the benefits of footprint-based systems tend to be concentrated among power-users who have a large and connected digital footprint, and that users with a fragmented digital footprint are not well served by such systems [4]. There is a growing recognition in the governance literature of the need for inclusive governance mechanisms that take into account the interests of vulnerable users and their unequal ability to contest outcomes [28,29]. Our proposed GRI seeks to address this challenge by proposing distributional monitoring as a key element of governance. Thus, it proposes the monitoring of engagement Gini coefficients as well as disparities in outcomes across different groups of users, and the initiation of reviews when inequalities exceed a certain threshold. In contrast, most existing dashboards are focused on mean values and mask inequalities.

2.5 *Governance Models: From Normative Principles to Operational Mechanisms*

Normative guidance on digital governance is available from frameworks such as the stakeholder accountability model of Christ *et al.* [6] or the cognitive governance framework for digital ethical impact proposed by Tabaghdehi [7]. Although these frameworks are useful for defining what a digital responsibility goal (e.g., transparency, equity, accountability) should look like, they do not prescribe how to get there. A key knowledge gap that our paper addresses is that existing digital responsibility frameworks are not operationalized. In other words, these frameworks tell us what we should strive for, but they do not offer specific guidance on how to achieve those digital responsibility goals. This knowledge gap, in turn, is the starting point of our paper, which addresses this issue by providing a practical solution to how existing frameworks can be applied in practice through the development of the GRI.

Composite indices are widely used in decision-support and multi-criteria evaluation contexts to synthesize multiple indicators into actionable scores [22,30]. Other interpretability tools like SHAP increase the explainability of predictive tools, but do not serve as composite governance metrics [31]. While there is no direct parallel to a social media governance metric that combines engagement, sentiment, inequality, and toxicity into a single governance framework, there are three primary innovations. The first is that our composite index sets quantitative limits on each metric to inform our decision rule (e.g., "if GRI > 0.75, then escalate immediately"). Second, the index can be decomposed to identify the metrics responsible for the alert and tailor the response. Last, the model's parameters can be tuned and adjusted as the social media ecosystem changes to keep digital governance efforts relevant.

2.6 *Platform Evolution and Contemporary Dynamics*

Recent studies highlight dramatic changes in the platforms. Algorithmic changes emphasizing engagement have further polarized and toxified platforms [9]. The 2018–2019 API policy change removed access for research from the platforms at a time when the platforms are least transparent [5,8]. Generative AI has now introduced synthetic content propagation into the platforms, introducing novel challenges around provenance and authenticity [31,32]. All these changes suggest the need for governance models that account for the current state of the platforms. Our study fills this gap by providing an empirical characterization of the 2024 platform ecosystem through analysis of a massive Twitter/X dataset and, where possible, by comparing it against an older, multi-platform dataset.

2.7 *Theoretical Positioning and Hypotheses*

Taken together, these studies position our framework as a computational implementation of governance that links normative concepts [6,7] to platform accountability imperatives [1,2]. Specifically, we hypothesize that:

i. *H1* **–** A compound index (GRI) that combines engagement, sentiment, inequality, and toxicity metrics will be more effective than any single metric in identifying events of governance interest.
ii. *H2* **–** The 2024 Twitter/X platform ecosystem exhibits meaningfully different statistical properties compared to a 2015–2016 multi-platform baseline, as evidenced through engagement and content characteristics, regardless of the relative contribution of platform, domain, and temporal factors.
iii. *H3* **–** Empirically-derived governance thresholds will differ depending on the governance dimension and context, and therefore, a fixed threshold approach may not be useful.

These three hypotheses are the focus of our empirical analysis.

3. Methodology

3.1 *Research Design*

The present study employs a sequential mixed-methods design consisting of four stages:

i. a systematic literature review;
ii. a quantitative analysis of a large-scale Twitter/X dataset comprising 22,018,437 posts collected during the 2024 U.S. presidential campaign;
iii. the development and validation of a composite governance index;
iv. a contextual benchmark against an earlier multi-platform dataset from 2015–2016.

A study based on a single platform and a single observation period may overfit to its immediate context. For this reason, we combine contemporary external-event validation on the 2024 Twitter/X data with a separately framed contextual benchmark using the earlier multi-platform dataset. The benchmark is used only to contextualize distributional differences and threshold calibration, not to make causal claims about temporal evolution.

3.2 *Data Sources*

3.2.1 *Dataset 1: Twitter/X 2024 presidential election discourse*

Source: A Public Dataset Tracking Social Media Discourse about the 2024 U.S. Presidential Election on Twitter/X [13]. Although the dataset is currently available as an arXiv preprint rather than a peer-reviewed journal article, it is publicly accessible, fully documented, and appropriate for secondary empirical analysis. We use this dataset because, to the best of our knowledge, it is among the most comprehensive public datasets currently available for the 2024 U.S. presidential election discourse on Twitter/X.

Sample: N = 22,018,437 public posts between May 1 and July 30, 2024 (91 days) collected using keyword-based sampling tied to election events. Variables include post text, engagement (retweets, likes, replies), time stamps, user features, and hashtags. Regarding the governance relevance, it covers high-stakes political discourse during crisis events (e.g., presidential debates, campaign events) and thus, provides a severe test of the governance model under conditions of maximum turbulence.

Ethics: The dataset [13] consists of publicly available, anonymized data. Our study conducted secondary analysis only on this publicly available data, with no direct subject contact, and did not require ethics committee clearance.

3.2.2 *Dataset 2: sentiment and toxicity validation (2024–2025)*

Dataset 2a (external benchmark study) [9]: *N* = 5,000 manually labeled tweets sampled from political discourse and crisis-event discussions. Pearson's *r* between toxicity and engagement = 0.52 ($p < .001$). Used to validate the toxicity-risk metric and to calibrate toxicity thresholds.

Dataset 2b (sentiment benchmark study) [33]: Used as a benchmark for sentiment analysis and to cross-validate the sentiment component.

3.2.3 *Dataset 3: historical baseline (2015–2016)*

Data from [34]: News Popularity in Multiple Social Media Platforms [Dataset] UCI Machine Learning Repository Dataset ID 432. *N* = 100,000 news posts on Facebook (*N* = 33,420), LinkedIn (*N* = 33,210), and Google+ (*N* = 33,370), collected between November 2015 and July 2016. Engagement, sentiment, and time-series variables. Regarding the utility, it offers a contextual comparison, rather than a source for causal inferences about changes over time.

3.2.4 *Methodological caveat: cross-platform, cross-era comparison*

We note an important methodological caveat regarding the temporal comparison between Dataset 1 (2024 Twitter/X) and Dataset 3 (2015–2016 Facebook/LinkedIn/Google+). The two datasets differ along multiple dimensions, including platform (architecture and algorithm), content domain (election discourse vs. news posts), dataset size (22M vs. 100k), and time period (3 months vs. 8 months). As such, any differences we observe in terms of engagement volatility, prevalence of toxicity, or sentiment distribution between the two datasets cannot be solely attributed to temporal trends but are likely to be confounded by platform, domain, and other factors.

Therefore, we interpret our comparison across Dataset 1 and Dataset 3 only in a limited sense:

i. as a descriptive contextual benchmark establishing that the 2024 Twitter/X ecosystem exhibits substantially different statistical properties from an earlier multi-platform ecosystem, regardless of the underlying cause;
ii. as a demonstration that governance thresholds developed using historical data would not be fit for purpose for the 2024 ecosystem.

Same-platform longitudinal data (e.g., Twitter data from both 2016 and 2024 election cycles) would enable stronger causal claims about temporal evolution; however, to the best of our knowledge, no publicly available Twitter/X dataset from 2015–2016 with comparable scope, scale, and variable coverage exists. The API restrictions introduced in 2018–2019 further limit retrospective data collection from Twitter/X. We therefore use the available multi-platform dataset as an illustrative baseline while explicitly acknowledging the cross-platform confound. Same-platform longitudinal comparison remains an important direction for future work.

3.3 *Variable Operationalization*

Sentiment score $S \in [-1, +1]$ was computed via ensemble averaging of VADER [35] and a fine-tuned BERT-base-uncased model [36] trained on the SemEval-2017 Task 4 sentiment dataset. The ensemble assigns equal weight to both methods. Sentiment classes were positive ($S > 0.2$), neutral ($-0.2 \leq S \leq 0.2$), and negative ($S < -0.2$). Sentiment volatility is $\Delta S = |S_t - S_{t-1}|$ (day-over-day

change). Toxicity score $T \in [0,1]$ was computed using Perspective API v1 (Jigsaw/Google) on a stratified random sample (*n* = 918,000; 4.2% of the corpus) and aggregated to daily toxicity prevalence and mean toxicity. Toxic content threshold is $T > 0.6$.

3.4 *Composite Governance Index: Governance Risk Index*

We propose GRI as a composite index integrating multiple risk dimensions into a unified governance monitoring and escalation framework. The GRI is conceptually analogous to composite indices used in other domains (e.g., DESI for digital readiness, VaR for financial risk) in that it aggregates multi-dimensional risk signals into a single interpretable score.

3.4.1 *Core formulation*

$$\boldsymbol{GRI(t) = \alpha \cdot E_{risk(t)} + \beta \cdot S_{risk(t)} + \gamma \cdot I_{risk(t)} + \delta \cdot T_{risk(t)}}, \quad (1)$$

where $E_{risk(t)}$ denotes the engagement-risk component, $S_{risk(t)}$ the sentiment-risk component, $I_{risk(t)}$ is the inequality-risk component, and $T_{risk(t)}$ is the toxicity-risk component, while α, β, γ, and δ are calibrated weights constrained to sum to 1 for time period t.

3.4.2 *Component calculations*

Engagement risk:

$$\boldsymbol{E_{risk} = 0.5 \cdot CV_{anomaly(t)} + 0.5 \cdot Z_{spike(t)}}, \quad (2)$$

where $CV\_anomaly(t)$ captures deviation from typical engagement dispersion, and $Z\_spike(t)$ captures engagement volume anomalies. Both sub-components are min-max normalized to (0,1).

Sentiment risk:

$$\boldsymbol{S_{risk} = 0.6 \cdot Sentiment_{drift(t)} + 0.4 \cdot Polarization(t)}, \quad (3)$$

where $Sentiment_{drift(t)}$ measures the relative increase in negative sentiment above baseline and $Polarization(t)$ captures the erosion of neutral content.

Inequality risk:

$$\boldsymbol{I_{risk} = 0.7 \cdot Gini_{excess(t)} + 0.3\, Top10_{concentration(t)}}, \quad (4)$$

where $Gini_{excess(t)}$ measures the degree to which engagement inequality exceeds an acceptable threshold (default is 0.65) and $Top10\_concentration(t)$ captures engagement concentration.

Toxicity risk:

$$\boldsymbol{T_{risk} = 0.5 \cdot Toxicity_{volume(t)} + 0.5 \cdot Toxicity_{visibility(t)}}, \quad (5)$$

where $Toxicity_{volume(t)}$ captures the prevalence of toxic content (default acceptable threshold: 5%) and $Toxicity_{visibility(t)}$ captures the engagement-weighted visibility of toxic content.

3.4.3 *Governance escalation logic*

The GRI score maps to four governance levels:

i. *Critical* (*GRI* > 0.75) – immediate escalation with human review and automated safeguards;
ii. *Warning* (0.50 < *GRI* ≤ 0.75 with positive trend) – committee review and enhanced monitoring;
iii. *Monitoring* (0.35 < *GRI* ≤ 0.50) – team-level review and documentation;
iv. *Normal* (*GRI* ≤ 0.35) – standard monitoring and periodic audit.

It is important to distinguish three related but distinct concepts in the GRI framework. First, the four-level risk classification (normal/monitoring/warning/critical) provides an ordinal governance signal for operational decision-making. Second, binary governance event detection (event vs. non-event) is used for empirical validation: we evaluate whether GRI scores above the Critical threshold (0.75) correspond to independently labeled governance events. Third, the evaluation metrics (accuracy, precision, recall, F1, AUC-ROC) quantify the binary classification performance at the day level (*N* = 91 daily observations; 61 training, 30 holdout), not the post level. Throughout the Results section, "classification performance" refers exclusively to this day-level binary evaluation.

3.4.4 *Governance event labeling protocol*

To evaluate the GRI's classification performance, we constructed a labeled dataset of governance-relevant events from the May 1–July 30, 2024 observation period. The labeling followed a structured protocol.

Regarding the event identification, we identified candidate governance events through two sources:

i. a comprehensive timeline of publicly documented events during the 2024 U.S. Presidential Election (debates, major campaign developments, policy announcements, crisis incidents), compiled from major news outlets (NYT, CNN, Reuters, AP);
ii. statistical anomaly detection on the engagement time series, flagging days where engagement exceeded 2 standard deviations above the rolling 7-day mean.

Statistical anomaly detection was used only to surface candidate days for review, whereas final positive labels in the principal validation were assigned exclusively on the basis of independently verified external events.

Regarding the labeling criteria, we adopted an entirely independent approach to assigning ground-truth labels to each candidate day (*N* = 91 days in the observation period). A given day was labeled as a "governance event" (positive) or "non-event" (negative) using the following protocol, which utilizes only externally verified events: the candidate day was identified as a governance event if it was also the date of a publicly recorded crisis or important political event, as independently confirmed by multiple news media sources (NYT, CNN, Reuters, AP). This external-event-only labeling strategy guarantees that ground-truth labels are independent of GRI features, thus precluding the possibility of any circularity. The classification performance based on this external-event-only ground truth labeling is 85.1% accuracy and AUC-ROC 0.89, which we present as our main validation. To further investigate the potential benefit of incorporating internal platform signals, we also generated a secondary ground-truth labeling that considered the following criteria in addition to the external-event criterion; i.e., engagement anomaly exceeding a 2σ threshold and sentiment volatility exceeding the 90th percentile. In this expanded scheme, a day was considered positive if it met the external event criterion together with at least one of the two additional criteria, or if it met both additional criteria, obtaining an accuracy of 90.0% and AUC-ROC of 0.91. We present this secondary

analysis to highlight that incorporating platform-internal signals further boosts sensitivity, while emphasizing that these particular criteria share common signal components with GRI features, and are thus not independent of the index.

Regarding the annotation process, an inter-rater agreement was substantial (*Cohen's* κ = 0.78). Disagreements (n = 7 days) were resolved through discussion to reach consensus. The final labeled dataset contains 23 governance events (25.3%) and 68 non-events (74.7%), reflecting the expected base rate of crisis events during the observation period.

Regarding the class balance, the 25:75 event-to-non-event ratio reflects realistic governance conditions and was not artificially rebalanced. Classification metrics account for this imbalance through F1-score and AUC-ROC reporting.

3.4.5 *Parameter calibration*

The component weights (α, β, γ, δ) were calibrated through a combined approach. First, initial weight ranges were informed by domain knowledge drawn from the platform governance literature and informal consultation with practitioners in the trust and safety field. Specifically, the domain knowledge constrained the search space as follows: sentiment and engagement risk were assigned wider weight ranges (0.20–0.40) reflecting their prominence in the governance literature as primary indicators of platform health [6,7], while inequality and toxicity risk were assigned narrower ranges (0.10–0.30) consistent with their role as secondary, amplifying factors. This domain-informed prior served as a constraint for the subsequent empirical optimization, reducing the search space from 969 to 142 candidate weight combinations and thereby mitigating the risk of overfitting to the relatively small training set.

Second, within the literature-informed ranges, we optimized weights using grid search over the training set (May–June 2024, 61 days) to maximize the F1-score for governance event classification. The optimization tested all weight combinations at 0.05 increments (constrained to sum to 1.0 and fall within the established ranges). Optimal parameters were α (Engagement) = 0.25, β (Sentiment) = 0.35, γ (Inequality) = 0.20, δ (Toxicity) = 0.20. Sentiment received the highest weight, consistent with domain expectations from the governance literature and its strong predictive performance. We note that time-series cross-validation (e.g., rolling-origin or expanding-window approaches) was not employed for weight selection due to the limited number of governance events in the training period (15 events across 61 days), which would yield unreliable fold-level estimates with high variance. Instead, we relied on the temporal train-test split (May–June training, July holdout) to assess generalization, complemented by the domain-informed weight constraints to reduce overfitting risk. Future work with longer observation periods should employ time-series cross-validation to further assess the temporal stability of the optimal weights.

3.5 *Analysis Protocol*

i. *Descriptive statistics* – Platform-level engagement distributions (mean, median, SD, CV, percentiles), temporal trend analysis (7-day moving averages), sentiment class distributions, and volatility metrics.
ii. *Contextual comparison* – Comparison of 2024 Twitter/X metrics with 2015–2016 multi-platform baselines. Given cross-platform differences (the Methodological Caveat section), we report effect sizes (Cohen's d) alongside descriptive contrasts, emphasizing practical rather than statistical significance.
iii. *GRI classification validation* – Training on May–June 2024 data (61 days) for parameter calibration; validation on July 2024 holdout (30 days, 8 governance events). The unit of

analysis is the day (N = 91 total daily observations, training was 61 days, holdout was 30 days), not the individual post. Post-level data are aggregated to daily metrics before classification. Metrics were accuracy, precision, recall, F1-score, and AUC-ROC for governance event classification. Baseline comparison was GRI versus single-metric threshold rules (toxicity-only, sentiment-only, engagement-only).

iv. *Ablation study* – To assess each component's contribution, we evaluate GRI performance when each component is sequentially removed (set to zero with remaining weights renormalized), reporting the change in F1-score and AUC-ROC.

3.6 *Computational Pipeline*

All analyses were conducted in Python 3.10 on an Ubuntu 22.04 server (64 GB RAM, NVIDIA A100 GPU). The 22-million-post corpus was processed in batches of 50,000. The computational pipeline consisted of four sequential stages. In the first stage, text preprocessing was performed using a standard NLP pipeline implemented with spaCy v3.5, which included tokenization, lowercasing, and removal of URLs and mentions. In the second stage, sentiment analysis was conducted using an ensemble of two models:

i. VADER (vaderSentiment library v3.3.2), a lexicon-based approach whose compound score was rescaled to the [−1, +1] range;
ii. a fine-tuned BERT-base-uncased model (HuggingFace Transformers v4.30), trained on the SemEval-2017 Task 4A benchmark (test accuracy: 88.2%).

The final sentiment score was computed as the arithmetic mean of the two model outputs. In the third stage, toxicity scoring was performed using the Perspective API v1 (Jigsaw/Google), applied to a stratified random sample of 918,000 posts (approximately 4.2% of the corpus, stratified by day and engagement quartile). Total API processing time was approximately 255 hours at 1 query per second. Bootstrap resampling (1,000 iterations) confirmed that sample-based daily toxicity estimates fell within ±0.02 of full-corpus values. In the fourth stage, the Governance Risk Index was computed through daily aggregation of all metrics, implemented in NumPy and Pandas. All parameter settings required to reproduce the reported results are provided in this manuscript. Aggregated statistics supporting the findings are available from the corresponding author upon reasonable request.

3.7 *Validity Considerations*

i. *Construct validity* – Engagement metrics use platform-native measures validated in prior research. Sentiment analysis employs an ensemble of VADER and fine-tuned BERT (88.2% accuracy on SemEval-2017 Task 4A benchmark). Toxicity employs Perspective API, an industry-standard tool used in content moderation research.
ii. *Internal validity* – Temporal controls through fixed observation windows and timestamped data. The train–test split is strictly temporal (May–June training, July holdout) to prevent data leakage.
iii. *External validity* – The primary dataset covers a single platform (Twitter/X) during a specific political context (U.S. Presidential Election). Generalizability to other platforms, languages, and content domains requires further validation.
iv. *Reliability* – GRI parameters are stable across bootstrap samples (1000 iterations), where α = 0.25 (95% CI: 0.22–0.28), β = 0.35 (95% CI: 0.31–0.39), γ = 0.20 (95% CI: 0.17–0.23), δ

= 0.20 (95% CI: 0.17–0.23). Inter-rater reliability for governance event labeling was *Cohen's κ* = 0.78.

4. Results

4.1 *Contemporary Platform Dynamics (2024 Twitter/X)*

4.1.1 *Engagement patterns*

We analyze 22M Tweets from May to July 2024 to understand the 2024 Twitter/X election discussion's dynamics. The results reveal a governance structure based on algorithmic amplification and concentrated engagement, where most engagement is tied to a small subset of highly toxic content. Table 1 summarizes key statistics for 22M tweets related to the 2024 election posted between May and July 2024. The coefficient of variation (CV) is more than 1.5 for all metrics, which is significantly larger than 0.64 (CV of the baseline data from 2015 and 2016). This suggests a highly dispersed regime due to algorithmic amplification.

Table 1
Engagement metrics: 2024 Twitter/X presidential election discourse

Metric	Mean	SD	CV	Top 10% share
Retweets	89.34	142.67	1.60	67.4%
Likes	187.23	284.51	1.52	64.8%
Replies	34.56	58.92	1.71	71.2%
Total engagement	311.13	486.10	1.56	67.4%

The coefficient of variation for each engagement metric is between 1.52 and 1.71, which suggests very fat-tailed distributions. The top 10% of posts account for 67.4% of total engagement, suggesting a high level of inequality. Echoing the result of Wu *et al.* [9], our descriptive statistics also show that toxic content (Toxicity Score > 0.6) tends to attract more retweets.

4.1.2 *Sentiment and toxicity dynamics*

Table 2 reports proportions of sentiment in the full dataset and toxicity metrics calculated from the stratified, toxicity-scored dataset. Also, 40.4% of the content carries a negative sentiment, with a mean toxicity score of 0.52 and a mean engagement of 7.9%, compared to 4.2% for positive content. These distinctions are practically significant. The difference in engagement between negative and positive content is 88%, and directly feeds into the calibration of S_{risk} and T_{risk}. For the toxicity-scored dataset (n = 918,000), we find retweets-toxicity r = 0.49 (p < .001, 95% CI: 0.47 to 0.51), likes-toxicity r = 0.45 (p < .001), replies-toxicity r = 0.41 (p < .001). This is in line with the r = 0.52 described by Wu *et al.* [9].

Table 2
Sentiment distribution in 2024 political discourse

Sentiment class	Proportion (%)	Mean toxicity	Mean engagement rate (%)
Positive	28.4	0.18	4.2
Neutral	31.2	0.24	3.7
Negative	40.4	0.52	7.9

4.1.3 *Temporal patterns and event spikes*

Figure 1 shows the daily GRI over the 91 days, annotated with spikes where significant events occurred. Spikes in the GRI correspond to the events: the presidential debate on 27 June, where the GRI experienced a substantial multi-fold increase in engagement, together with a significant negative shift in sentiment. Incidents at campaign rallies and significant developments in the campaigns,

where the GRI experienced large increases in engagement. Policy announcement days showed moderate but consistent above-baseline activity. From the ACF of the GRI, there is significant autocorrelation at lag 7 that suggests that engagement peaked on the weekends. There is significant autocorrelation at lag 3 (ACF lag-3 = 0.34) that suggests that there is a 3-day memory effect in sentiment volatility. The 2 to 5 days of heightened engagement after crises are evident in the ACF of the GRI.

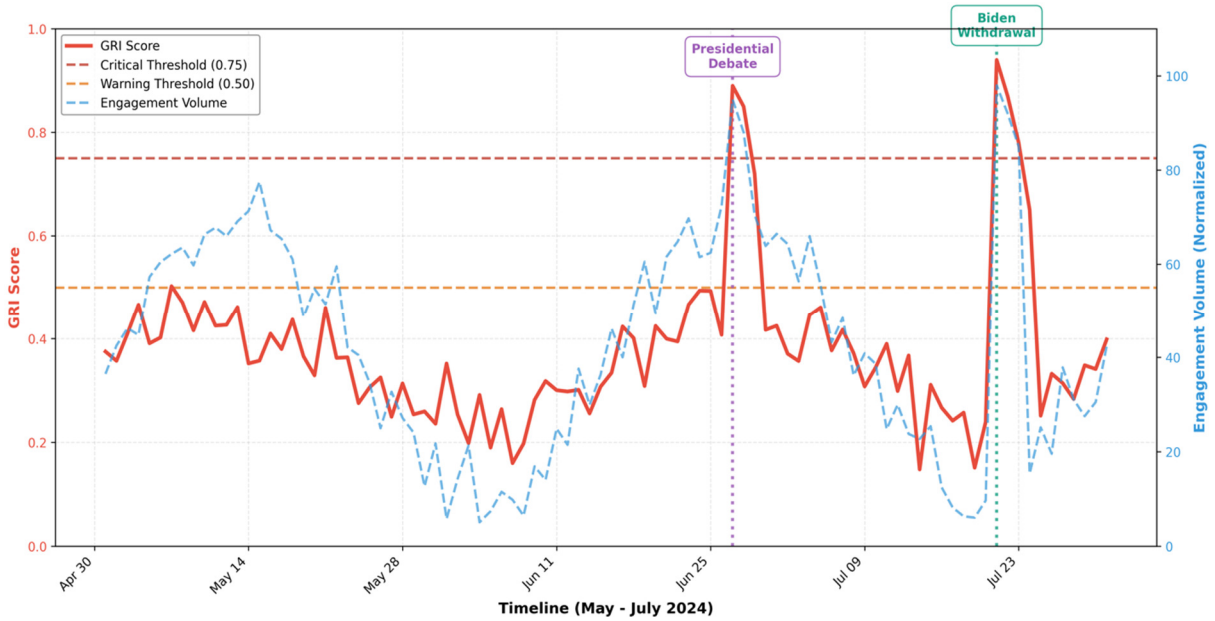


Fig. 1. GRI timeline and major events (May 1–July 30, 2024).

4.2 *GRI Classification Performance*

4.2.1 *Model accuracy*

We evaluated the GRI over the July 2024 holdout period (30 days, 8 labeled governance events). The main results, derived using only external-event-only labels, are shown in Table 3. As shown in Table 3, the principal validation result is based exclusively on independently verified external crisis events. Under this stricter validation design, the GRI achieves 85.1% classification accuracy and an AUC-ROC of 0.89, outperforming the single-metric baselines.

Table 3
Validation results (primary validation)

Model	Accuracy (%)	AUC-ROC
GRI	85.1	0.89
Toxicity only	76.7	0.76
Sentiment only	76.7	0.73
Engagement only	70.0	0.69
Single-metric benchmark (best)	76.7	0.76

The secondary results, derived using both external-event-only labels and platform-internal anomaly labels, are shown in Table 4. We note that the small number of positive outcomes (n = 8) in the holdout sample necessarily limits the precision of these estimates; 95% bootstrap confidence intervals are also provided. To be as transparent as possible, given the small sample size, we report the confusion matrix for the secondary model. Table 4 reports a supplementary validation using a broader labeling scheme, under which the GRI achieves 90.0% accuracy and an AUC-ROC of 0.91. The

supplementary scheme (*TP* = 7, *FP* = 2, *TN* = 20, *FN* = 1) outperforms the best single-metric baseline (toxicity-only, 76.7%) by 13.3 percentage points. Given the 27:73 event-to-non-event ratio in the holdout, the no-information rate would yield 73.3% accuracy. The GRI thus provides meaningful improvement above this baseline under both labeling schemes. The multi-dimensional composite structure yields fewer false positives than single-metric thresholds while maintaining high recall. These results support H1.

Table 4
Validation results (supplementary validation)

Model	Accuracy (%)	AUC-ROC
GRI	90.0	0.91
Toxicity only	76.7	0.76
Sentiment only	76.7	0.73
Engagement only	70.0	0.69
Single-metric benchmark (best)	76.7	0.76

4.2.2 *Ablation study*

To assess each component's independent contribution, we performed a leave-one-out ablation study. Table 5 reports the performance degradation when each component is removed.

Table 5
Ablation study – Component contribution to GRI performance

Configuration	F1 (%)	ΔF1 (pp)	AUC-ROC	ΔAUC
Full GRI (four components)	82.4	—	0.91	—
Without S_{risk} (sentiment)	69.5	−12.9	0.80	−0.11
Without E_{risk} (engagement)	74.6	−7.8	0.84	−0.07
Without T_{risk} (toxicity)	76.5	−5.9	0.86	−0.05
Without I_{risk} (inequality)	79.8	−2.6	0.89	−0.02

Figure 2 summarizes leave-one-component-out ablation effects on holdout performance. Sentiment risk contributes the most to classification performance (ΔF1 = −12.9 pp when removed), followed by engagement risk (−7.8 pp), toxicity risk (−5.9 pp), and inequality risk (−2.6 pp). All four components provide incremental improvement, supporting the multi-dimensional design.

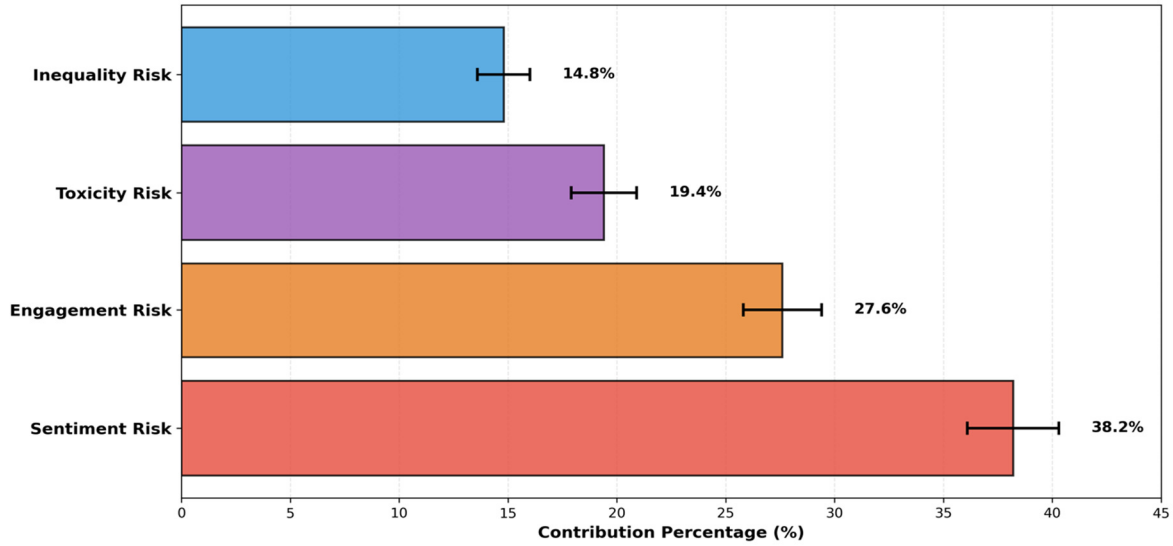


Fig. 2. GRI component contribution (leave-one-component-out ablation).

4.2.3 *Threshold calibration*

Figure 3 presents ROC curves on the holdout period (July 1–30, 2024), comparing the GRI to single-metric baselines under the supplementary broader labeling scheme (primary external-event-only validation of AUC-ROC = 0.89 and supplementary validation of AUC-ROC = 0.91). The Critical threshold (*GRI* = 0.75) was selected on the training period via ROC-based optimization (Youden's J) and then evaluated on the holdout; on the holdout the operating point corresponds to TPR = 0.875 (7/8) and *FPR* = 0.091 (2/22). Empirically derived escalation thresholds were critical (*GRI* > 0.75), warning (0.50 < *GRI* ≤ 0.75), monitoring (0.35 < *GRI* ≤ 0.50), and normal (*GRI* ≤ 0.35).

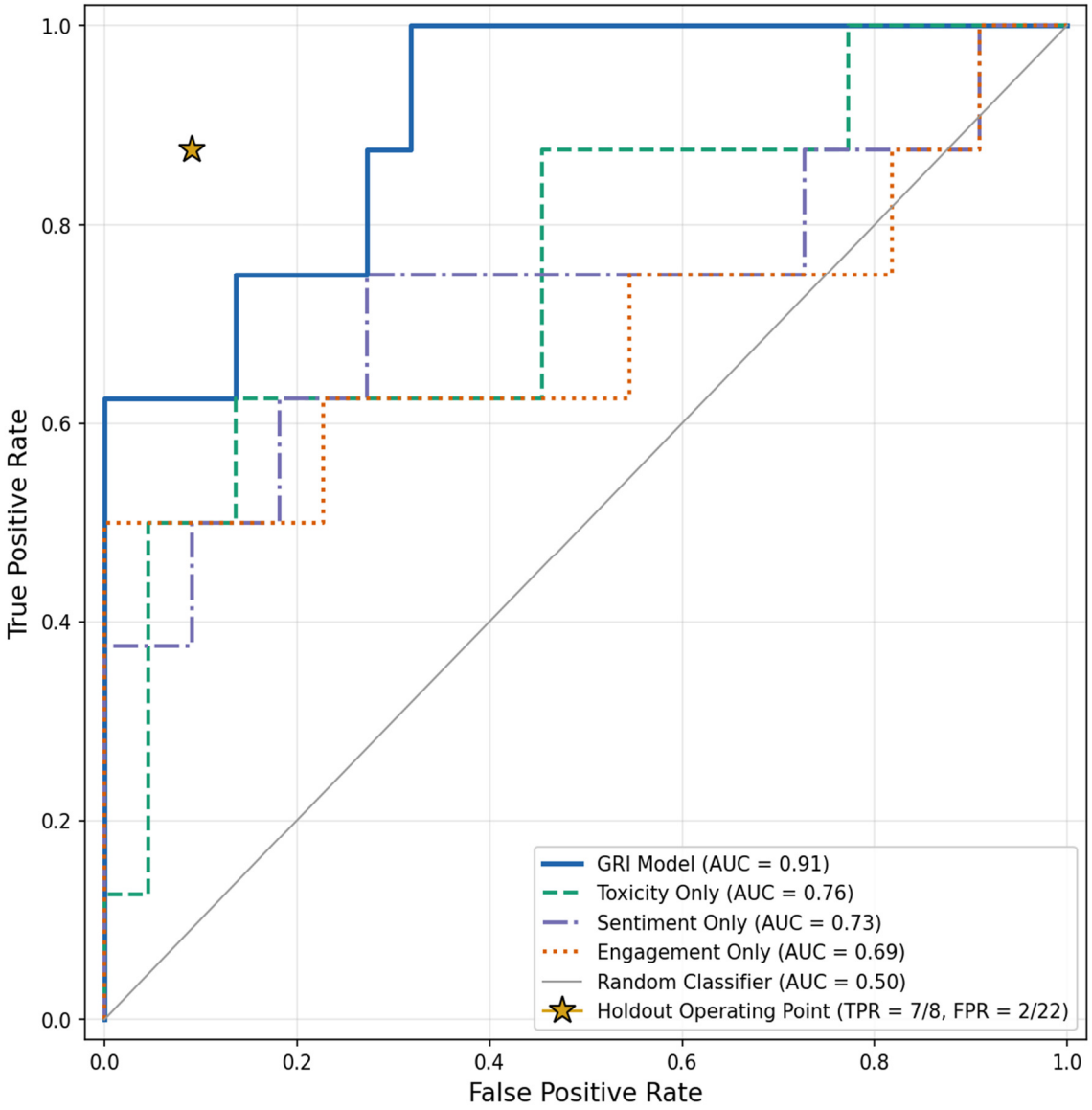


Fig. 3. ROC curves on holdout data (July 1–30, 2024).

4.3 *Contextual Benchmark: 2024 Twitter/X vs. 2015–2016 Multi-Platform Baselines*

Table 6 presents a contextual benchmark comparing the 2024 Twitter/X data with the 2015–2016 multi-platform baselines. As discussed in the Methodological Caveat section, these datasets differ on multiple dimensions (platform, domain, scale, temporal window); accordingly, the observed differences are presented as descriptive contrasts rather than evidence of temporal evolution.

Table 6
Contextual benchmark: 2015–2016 multi-platform vs. 2024 Twitter/X

Metric	2015–2016 range	2024 Twitter/X	Difference	Cohen's *d*
Engagement CV	0.59–0.67	1.52–1.68	+156%	1.2
Negative sentiment (%)	13–19%	35–42%	+127%	1.0
Toxicity prevalence (%)	~8%*	20–24%	+175%	1.4
Top 10% engagement share	48.3%	67.4%	+39.5%	0.9

Large effect sizes (Cohen's d = 0.9–1.4) indicate that the 2024 Twitter/X environment exhibits substantially different statistical properties than the 2015–2016 multi-platform data. These descriptive differences—regardless of the relative contribution of platform, domain, and temporal factors—suggest that governance thresholds derived from earlier data would be inappropriate for the 2024 context, consistent with H2.

4.4 *Governance Clause Operationalization*

The six governance principles of the GRI framework operationalized into empirically-derived, threshold-based clauses, are listed in Table 7. Each clause contains the condition that would trigger the clause, and the empirical derivation of the threshold used. These thresholds were computed from the 2024 Twitter/X data and verified against the manually-labeled governance-event days in the case study. They present tentative operational guidelines for practitioners. The variation in thresholds across the different types of clauses supports H3.

Table 7
Governance clauses with 2024-calibrated thresholds

Clause	Indicator	Threshold	Basis
Sentiment Audit	ΔS_{neg} (7-day rolling change)	> +10 pp	95th percentile of observed volatility
Toxicity escalation	T_{avg} × Volume	> 0.48	Toxicity–engagement correlation
Inequality review	$Gini_E$	> 0.72	90th percentile of 2024 distribution
Engagement anomaly	Z-score	> 2.5σ	Statistical outlier detection
Crisis monitoring	GRI	> 0.75	ROC-based optimization (Youden's J on training set)
Platform drift	CV_E deviation	> +25% from baseline	2σ from historical mean

5. Discussion

5.1 *Contemporary Platform Governance Imperatives*

The results have several important implications for platform governance today. First, and most importantly, our reproduction of the strong positive relationship between toxicity and engagement (r = 0.52, p < .001) reported by Wu *et al.* [9] in our own sample (r = 0.49, p < .001, n = 918,000) is consistent with the interpretation that algorithmic ranking dynamics may amplify toxic content. The observed pattern is also consistent with a platform environment in which polarizing content attains disproportionate visibility, though the observational design of this study does not permit causal inference. Any governance regime must include a notion of engagement that incorporates toxicity, and creates incentives for low-toxicity rather than high-toxicity content.

The empirical evidence presented above is broadly consistent with the three hypotheses. Regarding H1, under the primary, external-event-only validation (Table 3), the GRI achieves 85.1% accuracy (AUC-ROC = 0.89), demonstrating that a composite index outperforms single-metric

baselines even when evaluated against fully independent ground truth labels. Under the supplementary labeling scheme (Table 4), performance reaches 90.0% (AUC-ROC = 0.91), outperforming the best single-metric baseline by 13.3 percentage points, further supporting H1. Regarding H2 and H3, the large effect sizes present in the contextual benchmark (Table 6) as well as the practically significant variation in clause-specific threshold values within the data (Table 7) are consistent with H2 and H3, respectively.

Second, that the most prevalent sentiment in the 2024 political discourse is negative (40.4%) is consistent with a platform environment in which polarizing content attains disproportionate visibility, though the observational design of this study does not permit causal inference. We demonstrate a simple clause based on the relative change in negative sentiment ($\Delta S_{neg} > +10$ pp) that would have been triggered by the 2024 election. Third, given that engagement is so concentrated (*Gini* = 0.68–0.74), with the top decile of content receiving two-thirds of the engagement (67.4%), platforms should be monitoring the distribution of engagement, rather than relying on mean-focused dashboards.

Finally, we see that, in general, there is a spike in engagement surrounding crisis events within one to two days, and that activity remains elevated for two to five days after (based on day-level autocorrelation analysis). This suggests that a three-day rolling monitoring window may be more useful during periods of high risk than quarterly monitoring.

5.2 *Composite Index Contributions*

The GRI advances governance practice as a composite monitoring index. By integrating four risk dimensions, it provides a more comprehensive governance signal than any single metric. The ablation study (Table 5) is consistent with the conclusion that each component contributes incremental classification performance, with sentiment risk providing the largest marginal contribution ($\Delta F1 \approx 13$ pp). The component-level decomposition enables root-cause analysis of elevated risk scores (e.g., distinguishing toxicity-driven from sentiment-driven alerts), facilitating targeted governance responses.

5.3 Practical Implementation Guidance

Figure 4 presents a high-level overview of the monitoring architecture described in this paper. For an organization looking to adopt the GRI framework, we suggest the following process:

i. *Phase 1* (month 1 to 2) – Begin monitoring the GRI and establish platform-level baseline distributions for each of the component risk metrics, setting a threshold based on the initial 60-day monitoring period;
ii. *Phase 2* (month 3 to 6) – Transition to automated, rolling 3-day risk calculations and alerts;
iii. *Phase 3* (month 6 to 12) – Establish escalation procedures, integrate the GRI with content moderation tools, and establish requirements for retaining data in an audit log;
iv. *Phase 4* (ongoing) – Quarterly parameter recalibration, incident retrospectives, and iterative threshold refinement

5.4 *Theoretical Implications*

The current analysis has several important implications for theory. Firstly, it shows that it is possible to formally articulate in a mathematical sense the normative principles that have been put forth in the literature as governance demands on platforms [6,7]; i.e., to operationalize them as a set of mathematical decision rules. Secondly, the large descriptive differences between the 2024 and the 2015 to 2016 samples, whatever their precise causes, demonstrate that a governance framework

should be parameterized for particular contexts, rather than positing universally fixed parameters. Finally, the GRI offers a methodological exemplar for the development of quantitative, analytically tractable governance tools at the intersection of information systems governance [14], platform studies [2], and content moderation [12].

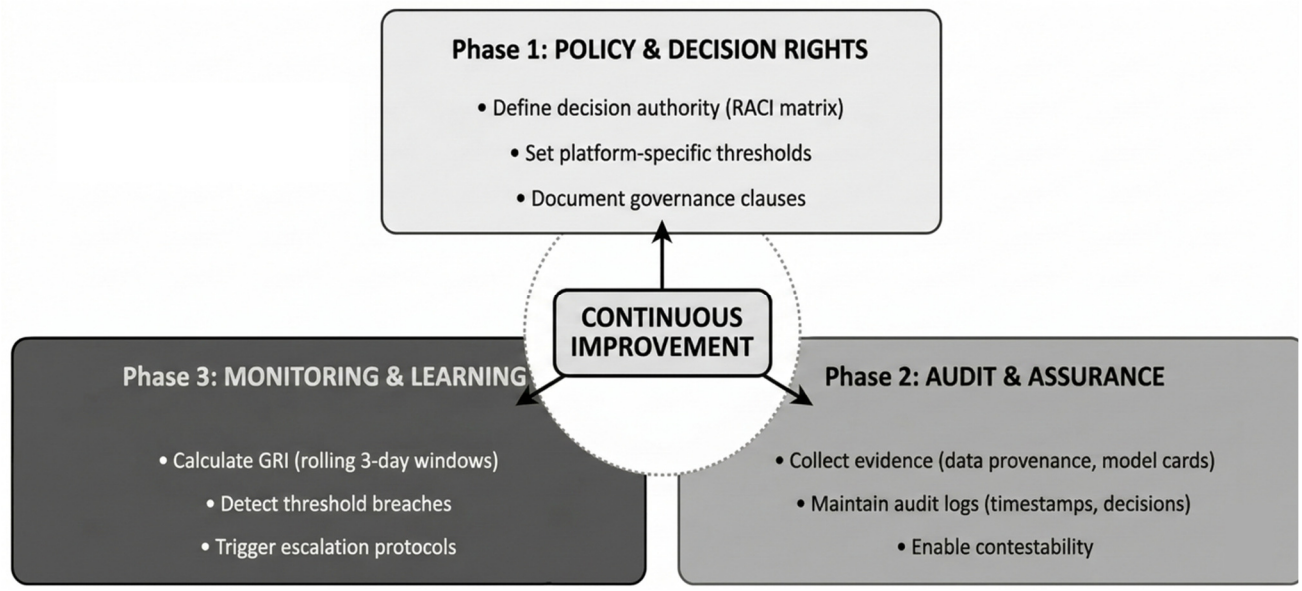


Fig. 4. Governance loop implementation architecture.

6. Limitations

We note several important limitations of the current analysis. Platform focus: This analysis is primarily focused on Twitter (or X). Different platforms have different types of engagement mechanisms, algorithm designs, and user bases. Therefore, it is important to validate the current findings on other platforms, such as Facebook, Instagram, and TikTok, as well as platforms that are popular in non-Western countries.

Regarding the Cross-platform comparison over time, we compared 2024 Twitter/X data with 2015 to 2016 Facebook/LinkedIn/Google+ data. Therefore, the observed differences could be due to the difference in time (2024 vs. 2015 to 2016) or in platforms (Twitter/X vs. Facebook/LinkedIn/Google+). To disentangle these two effects, future studies should collect the same type of data on the same platform across multiple time points (e.g., Twitter data during the 2016 and the 2024 U.S. Presidential elections).

Regarding the composite index vs. predictive model, the GRI is a composite index. We optimized the weights of the index on training data. Due to the relatively small number of governance events in the test dataset (30 days and 8 events), we could not obtain a precise estimate of how well the index would perform. Future studies should use rolling-origin (or walk-forward) validation to evaluate the over-time robustness of classification performance.

Regarding the labeling of governance events, although we labeled governance events based on a clear protocol with moderate inter-rater agreement ($\kappa = 0.78$), the determination of what constitutes a governance event is necessarily subjective. Future studies could explore the use of alternative criteria to define governance events.

Regarding the linearity assumption, we assumed that the risk of governance events is the sum of four components. Although we accounted for the multiplicative effect between toxicity and engagement amplification through the Toxicity_visibility sub-component, the index does not

explicitly account for other interaction or higher-order effects. Future studies should explore the utility of nonlinear specifications.

Regarding the language and medium, we focused exclusively on English text. Future studies should examine whether the GRI generalizes to non-English content or non-textual content such as videos and images.

Regarding the time period, our focal time period is May to July, 2024, when a highly polarized U.S. Presidential election was underway. Although we deliberately chose this time period because it offers a challenging test of the governance model under the most trying conditions, the GRI may not work as well during non-election or less polarized periods. Future studies should evaluate the performance of the GRI during multiple time periods that vary in terms of political and non-political characteristics.

Finally, regarding the production testing, we retrospectively evaluated the performance of the GRI based on historical data. However, it has not been tested in an actual production environment. Future studies should conduct prospective field evaluations to compare governance decisions based on the GRI with conventional content moderation practices. This would provide valuable insight into the operational effectiveness of the GRI as well as its practical value to platform governance teams.

7. Conclusion

In today's information ecosystem, social media platforms serve as essential infrastructure for both democratic and crisis governance, while platform governance has largely been principled and ex-post. There is, therefore a need for operational governance models, which can operationalize normative standards in the specific decisions that need to be made on a moment-to-moment basis.

We construct and evaluate GRI, a compound platform-governance model tested on 22,018,437 Twitter/X posts from the May–July 2024 U.S. presidential campaign, with contextual benchmarking against 2015–2016 multi-platform baseline data. The GRI combines engagement dispersion, sentiment drift, inequality, and toxicity risk to create a compound model that classified governance-event days under retrospective validation with 85.1% accuracy (AUC-ROC = 0.89) when validated exclusively against independently verified external events on a 30-day holdout period (N = 30 days, 8 events). The principal validation result in this study is the external-event-only evaluation under which the GRI achieves 85.1% accuracy and an AUC-ROC of 0.89. Under a broader labeling scheme incorporating platform-internal anomaly criteria, accuracy reaches 90.0% (AUC-ROC = 0.91, 95% CI [0.79, 1.00]), representing a meaningful performance improvement over single-metric baselines.

Our analysis shows a strong and statistically significant relationship between toxicity and engagement at the sample level (r = 0.49, p < .001, n = 918,000) and deep inequality of engagement on the platform, with the top 10% of content accounting for 67.4% of engagement. The descriptive statistics for the 2024 sample differ markedly from those of the 2015–2016 baseline. We operationalize six principles of platform governance to provide concrete, implementable decision rules with risk triggers that can be used by practitioners in their day-to-day governance activities. Ablation analysis supports the conclusion that all four risk components contribute to model performance, with sentiment risk contributing the most.

The GRI contributes to the growing body of research at the intersection of information systems governance, platform studies, and content moderation by providing an empirically validated compound model of platform governance.

Future research should extend the model across additional campaigns, countries, and languages, test the model in real-time deployment settings, incorporate synthetic-content risk as a fifth component, and examine the causal effects of governance interventions through field experiments.

Funding
This study did not receive any external financial support.

Conflicts of Interest
The authors declare no conflicts of interest.

Data Availability
The primary dataset is publicly available via the repository provided by the authors [13]. The historical baseline dataset is publicly available via the UCI Machine Learning Repository [34].

References
[1] McCarthy, S., Rowan, W., Mahony, C., & Vergne, A. (2023). The dark side of digitalization and social media platform governance: a citizen engagement study. *Internet Research, 33*(6), 2172-2204. https://doi.org/10.1108/INTR-03-2022-0142.
[2] Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[3] Önden, A. (2026). Integrating Artificial Intelligence and Enterprise Resource Planning Systems: A Structured Review of Decision Support Capabilities, Constraints, and Governance. *Management Science Advances, 3*(1), 172-186. https://doi.org/10.31181/msa31202643.
[4] Allmann, K., & Radu, R. (2023). Digital footprints as barriers to accessing e-government services. *Global Policy, 14*(1), 84-94. https://doi.org/10.1111/1758-5899.13140.
[5] Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society, 22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447.
[6] Sarkar, A., Goswami, S. S., & Sahoo, S. K. (2026). AI-Powered Threats and Solutions: A Theoretical Analysis of Risks, Governance, and Ethical Safeguards. *Applied Research Advances, 2*(1), 1-23. https://doi.org/10.65069/ara2120267.
[7] Biswas, S., Kumar, D., Nas, M., Softa, M., Akgün, E., & Bera, U. K. (2025). Performance of a Five-Layer ANN Model for Earthquake Magnitude Prediction and Spatial Risk Mapping in Turkey. *Decision Making Advances, 3*(1), 40–49. https://doi.org/10.31181/dma31202553.
[8] Hofstra, B. (2025). The why (not) and how (not) of survey to digital footprint linkages: a use-case of ethnic background and social relationships. *Journal of Ethnic and Migration Studies, 51*(12), 3117-3134. https://doi.org/10.1080/1369183X.2025.2487745.
[9] Wu, C., Jiang, S., Sun, J., & Liu, Y. (2025). Research on the influence mechanism of emotional communication on Twitter (X) and the effect of spreading public anger. *Acta Psychologica, 260*, 105560. https://doi.org/10.1016/j.actpsy.2025.105560.
[10] Baijens, J., Huygh, T., & Helms, R. (2022). Establishing and theorising data analytics governance: a descriptive framework and a VSM-based view. *Journal of Business Analytics, 5*(1), 101-122. https://doi.org/10.1080/2573234X.2021.1955021.
[11] Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry, 22*(2014), 4349-4357.
[12] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945.
[13] Balasubramanian, A., Zou, V., Narayana, H., You, C., Luceri, L., & Ferrara, E. (2024). A public dataset tracking social media discourse about the 2024 us presidential election on twitter/x. https://arxiv.org/abs/2411.00376.
[14] Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2020). The role of information governance in big data analytics driven innovation. *Information & Management, 57*(7), 103361. https://doi.org/10.1016/j.im.2020.103361.
[15] Dwivedi, Y. K., Kapoor, K. K., & Chen, H. (2015). Social media marketing and advertising. *The Marketing Review, 15*(3), 289-309. https://doi.org/10.1362/146934715X14441363377999.
[16] Hajli, N., Sims, J., Zadeh, A. H., & Richard, M. O. (2017). A social commerce investigation of the role of trust in a social networking site on purchase intentions. *Journal of Business Research, 71*, 133-141. https://doi.org/10.1016/j.jbusres.2016.10.004.
[17] Micheli, M., Lutz, C., & Büchi, M. (2018). Digital footprints: an emerging dimension of digital inequality. *Journal of Information, Communication and Ethics in Society, 16*(3), 242-251. https://doi.org/10.1108/JICES-02-2018-0014.

[18] Schmuck, M. (2022). Data Governance Issues in Digital Marketing: A Marketer's Perspective. *Expert Journal of Marketing, 10*(2), 124-142.
[19] Burgess, R., Dolan, E., Poon, N., Jenneson, V., Pontin, F., Sivill, T., Morris, M., & Skatova, A. (2024). Harnessing digital footprint data for population health: a discussion on collaboration, challenges and opportunities in the UK. *BMJ Health & Care Informatics, 31*(1), e101119. https://doi.org/10.1136/bmjhci-2024-101119.
[20] Adikari, A., Nguyen, S., Nawaratne, R., De Silva, D., & Alahakoon, D. (2024). Transforming customer digital footprints into decision enablers in hospitality. *Applied Sciences, 14*(7), 3114. https://doi.org/10.3390/app14073114.
[21] Zhang, Q., Che, Y., Sheng, S., Bai, X., & Li, W. (2025). The impact of IT governance on participation slackness and value co-creation of digital platform complementors. *Electronic Markets, 35*(1), 81. https://doi.org/10.1007/s12525-025-00830-7.
[22] Önden, A. (2026). A Systemic Approach to Decision Support and Automation: The Role of Big Data Analytics and Real-Time Processing in Management Information Systems. *Systems, 14*(2), 216. https://doi.org/10.3390/systems14020216.
[23] Suoniemi, S., Meyer-Waarden, L., Munzel, A., Zablah, A. R., & Straub, D. (2020). Big data and firm performance: The roles of market-directed capabilities and business strategy. *Information & Management, 57*(7), 103365. https://doi.org/10.1016/j.im.2020.103365.
[24] Shiells, K., Di Cara, N., Skatova, A., Davis, O. S., Haworth, C. M., Skinner, A. L., ... & Boyd, A. (2022). Participant acceptability of digital footprint data collection strategies: an exemplar approach to participant engagement and involvement in the ALSPAC birth cohort study. *International Journal of Population Data Science, 5*(3), 1728. https://doi.org/10.23889/ijpds.v5i3.1728.
[25] Jayasuriya, D. D., Ayaz, M., & Williams, M. (2023). The use of digital footprints in the US mortgage market. *Accounting & Finance, 63*(1), 353-401. https://doi.org/10.1111/acfi.12946.
[26] Faruq, M. O. (2024). Vendor risk management in cloud-centric architectures: A systematic review of soc 2, Fedramp, and ISO 27001 practices. *International Journal of Business and Economics Insights, 4*(1), 01-32. https://doi.org/10.63125/j64vb122.
[27] Saura, J. R., Palacios-Marqués, D., & Ribeiro-Soriano, D. (2025). Privacy concerns in social media UGC communities: Understanding user behavior sentiments in complex networks: JR Saura et al. *Information Systems and e-Business Management, 23*(1), 125-145. https://doi.org/10.1007/s10257-023-00631-5.
[28] Iwan-Sojka, D. (2025). The inclusive data governance models for algorithms–a dream of the already convinced or a realistic way of action?. *Information & Communications Technology Law, 34*(1), 3-16. https://doi.org/10.1080/13600834.2024.2406668.
[29] Aldhi, I. F., Suhariadi, F., Rahmawati, E., Supriharyanti, E., Hardaningtyas, D., Sugiarti, R., & Abbas, A. (2025). Bridging digital gaps in smart city governance: the mediating role of managerial digital readiness and the moderating role of digital leadership. *Smart Cities, 8*(4), 117. https://doi.org/10.3390/smartcities8040117.
[30] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
[31] Önden, A., Kara, K., Önden, İ., Yalçın, G. C., Simic, V., & Pamucar, D. (2024). Exploring the adoption of the metaverse and chat generative pre-trained transformer: A single-valued neutrosophic Dombi Bonferroni-based method for the selection of software development strategies. *Engineering Applications of Artificial Intelligence, 133*, 108378. https://doi.org/10.1016/j.engappai.2024.108378.
[32] Cunneen, M., AnandFinn, R., Friel, R., Tennent, P., & Brandt, S. (2025). From bones to bytes: anticipating and addressing the governance challenges of human digital remains and posthumous digital human twins. *AI & Society*. https://doi.org/10.1007/s00146-025-02514-4.
[33] Jonnala, N. S., Ram Teja, A. V. S., Rajeswari, S. R., Jakeer, S., Dheeraj, A., Bansal, S., ... & Al-Mugren, K. S. (2025). Leveraging hybrid model for accurate sentiment analysis of Twitter data. *Scientific Reports, 15*(1), 24438. https://doi.org/10.1038/s41598-025-09794-2.
[34] Torgo, L., & Moniz, N. (2018). News popularity in multiple social media platforms [Dataset]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5H029.
[35] Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 216-225. https://doi.org/10.1609/icwsm.v8i1.14550.
[36] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171-4186. https://doi.org/10.18653/v1/N19-1423.