Predicting Social Unrest and Political Risk Using GDELT: A Systematic Literature Review

Abstract

This systematic review examines 48 studies (2013–2025) that leverage GDELT data to forecast social unrest and assess political risk. We trace the evolution of methodologies from early statistical models to graph-based, machine learning, and multisource early warning systems that integrate news, social media, and economic indicators. Thematic analysis highlights five areas: (1) media bias and representation challenges; (2) validation against curated datasets; (3) predictive modeling using GDELT features; (4) construction of unrest indices; and (5) hybrid frameworks. GDELT enables the detection of early unrest patterns and signals with a precision of 70–85% but faces issues such as regional bias, geolocation errors, and limited generalizability. We identify eight key research gaps and call for more robust, ethical, and transparent early warning systems to inform preventive policy and interdisciplinary applications in computational social science.

Keywords: social unrest, political risk, GDELT, early warning, event data, machine learning

1 Introduction

Social unrest - sustained through protests, demonstrations, riots, strikes, and other collective actions - poses serious risks to political stability, economic development, and social cohesion [1, 2]. Traditional forecasting approaches have relied on expert judgment and low-frequency indicators, often lacking the timeliness and resolution needed to effectively anticipate unrest [3, 4]. The emergence of large-scale event datasets like GDELT, ICEWS, and ACLED has transformed this landscape, offering near-continuous structured records of socio-political events from global news sources [5, 6]. GDELT is known for its extensive reach, tracking over 60,000 sources in more than 100 languages. It codes events based on location, actors (using CAMEO), tone, and time, with updates every 15 minutes. [5, 7]. This real-time stream has inspired a growing body of research aimed at using media-derived signals to predict unrest and assess political risk [8–11].

However, predictive use of GDELT is not without challenges. It inherits media biases—framing, regional disparities, and language issues—that distort coverage [1, 12–14]. Automated coding introduces errors such as duplication, misclassifications,

and geolocation imprecision [15–17]. Modeling temporal and spatial dependencies is essential for understanding complex social phenomena[18]. In response, researchers have adopted a variety of approaches: statistical models for latent dynamics [8, 19], graph and sequence models for cascades [9, 20, 21], machine learning classifiers [22–24], and multi-source systems that combine GDELT with social, economic, and network data [10, 11, 25].

This review examines 48 significant studies that use GDELT to predict political unrest. It outlines methodological trends, assesses the strengths and limitations of GDELT data, synthesizes theoretical insights, and highlights future research priorities.

2 Systematic Methodology

We conducted comprehensive literature searches to ensure a robust and representative review of the field. Our search strategy spanned major academic databases, including IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and ScienceDirect. Boolean queries were formulated by integrating keywords such as "GDELT," "social unrest," "protest prediction," "conflict early warning," and "event data." The temporal scope of our search was defined from January 2013, coinciding with GDELT's public release, through March 2025 to capture the most recent advances. To ensure interdisciplinary breadth, we also reviewed proceedings from leading conferences (such as ICWSM, KDD, and WebSci), relevant policy reports, and selected doctoral theses. In addition, we employed snowball sampling, systematically examining the reference lists of key papers to identify further relevant studies that may not have surfaced in initial database queries.

Our inclusion and exclusion criteria were designed to focus the review on studies making substantial use of GDELT data for empirical analysis or prediction of collective political actions. Specifically, we include works that explicitly utilized GDELT to analyze or forecast events such as protests, riots, strikes, or broader instability phenomena such as conflict outbreaks and coups. Eligible studies were required to report quantitative evaluation metrics such as precision, correlation coefficients, or area under the curve (AUC), and to offer novel methodological contributions or significant analytical findings. Studies were excluded if they were limited to technical descriptions of the GDELT platform, presented theoretical arguments without direct application to GDELT data, or relied exclusively on alternative event datasets without incorporating GDELT.

For each study we reviewed, we systematically recorded the central research questions and specific data sources to emphasize which GDELT features were used and whether supplemental datasets (such as social media or economic indicators) were included. We also documented the methodological frameworks employed, ranging from hidden Markov models and statistical or econometric techniques to graph mining and machine learning approaches. Additional fields captured included the spatial and temporal scope of the analysis, key empirical results and performance metrics, limitations noted by the authors, and any proposed directions for future research. These detailed records facilitated cross-study synthesis and allowed us to identify both convergent findings and areas of methodological or substantive divergence.

To organize the literature for synthesis and comparative analysis, we applied inductive coding to study abstracts and methodological sections, grouping studies into five primary thematic clusters. These themes are; media coverage biases and event representation issues; validation and comparison of GDELT data against other datasets; predictive modeling approaches using GDELT features; construction of unrest indices and analytical applications; and multisource and advanced forecasting frameworks that integrate GDELT with other data streams. This thematic classification is designed to balance methodological taxonomy with substantive research focus, providing a coherent narrative structure for the review and facilitating nuanced comparative analysis from both within and across thematic areas.

3 Media Coverage Biases and Event Representation Issues

Biases in media coverage and event representation pose significant challenges when using GDELT and similar large-scale event datasets in computational social science research on social unrest. These biases emerge through interconnected mechanisms that influence the selection and interpretation of events, ultimately affecting the analysis of the reliability and validity of political instability.

Agenda-setting and framing effects represent primary sources of bias in media-derived event data. Wanta et al. [12] demonstrated that the volume and tone of international news coverage significantly influence public perceptions, and negative coverage exerts disproportionate effects. This indicates that GDELT's event counts and tone metrics reflect editorial priorities and systematic news production biases alongside objective event characteristics. McLeod [3] identified the "protest paradigm" in which mainstream media systematically emphasize violence and disruption in protest coverage, relying heavily on official sources, and marginalize protester perspectives. This paradigm delegitimizes dissent and leads automated coding systems to undercount peaceful demonstrations or mischaracterize events based on sensationalist language.

Regional and linguistic biases compound these framing effects through systematic coverage disparities. Kwak and An [7] found that only approximately 25% of the variance in global disaster coverage within GDELT can be explained by objective factors, the remainder attributable to strong regional biases favoring media-rich areas such as Europe and North America. Despite GDELT's expansion to include more diverse sources, English-language and Western media continue to dominate, skewing event representation and underrepresenting unrest in less-covered regions [13]. Barrett et al. [1] highlighted that media-based indices are limited by the underlying media landscape, often reflecting media attention rather than the true scope or intensity of the unrest.

There are technical challenges in automated extraction of events that introduce additional bias and noise layers. Sjovaag and Stavelin [15] documented methodological difficulties in coding online news, including high update frequencies, widespread duplication, and inconsistent metadata structures. Hammond and Weidmann [26]

revealed that GDELT's automated geolocation algorithms exhibit systematic capital bias, clustering events near urban centers while underrepresenting rural unrest. Their comparison with hand-coded datasets showed weak spatial correlation (0.20-0.26) despite stronger temporal correlation (0.33-0.64), indicating problematic spatial accuracy.

The limitations of the automated coding system further exacerbate representational challenges. Schrodt and Van Brackle [16] noted that even optimized machine coding systems achieve only about 75% accuracy, with errors arising from ambiguous language, complex sentence structures, and translation inaccuracies. Event ontologies necessarily simplify complex political phenomena into predefined categories, potentially losing crucial nuance and context. The selective nature of news reporting makes datasets inherently systematic, while current systems struggle with non-English languages and fail to distinguish between major and minor events effectively.

Predictive modeling approaches using GDELT data must account for these systematic biases. Qiao et al. [8] acknowledged that the sole reliance of their Hidden Markov Model on GDELT may introduce biases due to coverage inconsistencies. Studies employing graph-based approaches for crisis detection face challenges from dataset inconsistencies when combining GDELT with other sources [9]. Gao et al. [27] acknowledged that disparities in media coverage lead to overrepresentation of high-profile conflicts while underreporting localized crises.

These media-driven priorities create systematic spatial, thematic, and temporal distortions in GDELT event data that researchers must critically consider. Event coding inherits journalists' framing choices, rapid news cycles challenge deduplication efforts, and automated processing introduces systematic biases, necessitating methodological advances in bias correction and critical understanding of social and political forces shaping global news flows.

4 Validation and Comparison of GDELT Data

Comparison of GDELT with human-coded and other event datasets highlights ongoing issues with spatial accuracy, event categorization, and overall dependability. Hammond and Weidmann [26] performed one of the initial micro-level assessments by contrasting GDELT's machine-coded geolocations with hand-coded records from ACLED and GED for FARC-related events in Colombia between 2002 and 2009. Using spatial log-linear regression and error-variance mapping, they demonstrated that GDELT systematically clusters events near capitals and major urban centers, resulting in weak spatial correlations ($/rho = 0.26$ with ACLED; $/rho = 0.20$ with GED) despite moderate temporal alignment ($/rho = 0.64$; $/rho = 0.33$). This "capital-bias" geolocation error can displace true conflict hotspots by tens of kilometers, undermining microlevel analyses of rural unrest.

Automated event classification adds another layer of noise. Schrodt and Van Brackle [16] audited a random sample of GDELT's CAMEO CODED protest and violence events against human annotations, finding only 75% agreement. Errors arise from complex sentence structures, limited actor dictionaries, and mistranslations in non-English sources, leading to false positives (for example, economic protests flagged

as riots) and false negatives (peaceful demonstrations omitted). Such misclassifications distort frequency counts and downstream modeling efforts.

Compared to other large-scale news aggregators, GDELT's trade-off between volume and precision becomes apparent. Kwak and An [13] contrasted GDELT (covering 64 languages) with EventRegistry (14 languages), mapping daily country-level article counts. Although GDELT outpaces EventRegistry in raw volume, both exhibit high correlation in national coverage patterns; however, GDELT reports up to 3× more duplicate and noise-laden records per country per day.

Temporal validation through case studies underscores the mixed performance of GDELT. Keertipati et al. [28] employed change point detection in Sri Lankan Civil War and 2006 Fijian coup time series, showing that major regime changes align with abrupt jumps in GDELT counts, even as peripheral skirmishes remain underreported. Yonamine [29] correlated the GDELT-recorded protests with the Tel Aviv stock market returns, finding that significant protests preceded market declines by 3 to 5 days, illustrating the potential of the dataset for non-political temporal validation.

More recently, Raleigh et al. [6] demonstrated that differences in source selection and coding heuristics across datasets can obscure true patterns of political instability, cautioning that automated emphases on GDELT volume can inflate apparent unrest in media-rich countries.

To mitigate these limitations, best practices now include: (1) rigorous duplicate-removal and bounding-box filters for geoparsing; (2) hybrid human-machine validation loops to refine actor dictionaries; and (3) triangulation with hand-coded or alternative NLP pipelines. This preprocessing is critical for enhancing spatial, temporal, and thematic fidelity in any GDELT-based unrest analysis.

5 Predictive Modeling Approaches Using GDELT

Early work in statistical and econometric modeling demonstrated that GDELT event counts could serve as effective predictors of unrest. Qiao et al. [8] introduced a Hidden Markov Model (HMM) that treats verbal and material conflict counts from GDELT as emissions generated by latent tension states. By defining three hidden states—low tension, rising tension, and high tension—and estimating transition probabilities through the Baum–Welch algorithm, they captured typical escalation sequences (for example, government threats → mass protests). In balanced accuracy, their HMM outperformed logistic regression baselines by 7% to 27%, particularly improving recall for high-tension episodes in Southeast Asian case studies. Likewise, Gao et al. [27] demonstrated that unsupervised information-theoretic indicators, specifically spikes in relative entropy of event-type distributions and sudden breakdowns in pairwise correlations, preceded major crises such as the Arab Spring. By computing the daily entropy on the CAMEO categories and tracking divergence from historical correlation structures, they showed that macro-level unpredictability in GDELT streams can foreshadow political upheaval without requiring labeled training data.

Network-aware methods further enriched predictive performance by leveraging relational structures among actors. Keneshloo et al. [9] treated the GDELT dataset as a dynamic interaction graph in which nodes represent actors and weighted edges

represent event frequencies (e.g., protester-state confrontations). They extracted frequent subgraphs using a gSpan-inspired algorithm that distinguished crisis months from non-crisis months in five Latin American countries. Incorporating these graph-based features into an SVM classifier significantly improved F1-scores over count-only models, underscoring the importance of capturing recurrent interaction motifs (such as state coercion chains) as early warnings of unrest.

More recent work has applied high-dimensional machine learning classifiers to GDELT features. Zebrowski and Afli (2024) [24] extracted over 200 features, including lagged event counts, sentiment shifts (AvgTone derivatives), spatial dispersion metrics, and clustering coefficients, to train Random Forests and Bayesian neural networks for country-month instability forecasting. The Random Forest achieved roughly 85% contemporaneous accuracy and 75% next-month accuracy, outperforming SVMs and KNN models. Feature importance analyses highlighted that lagged spikes in protest and riot counts, together with sudden sentiment reversals, were the strongest predictors. Ozdemir (2018) [30] similarly used Random Forest regression on AvgTone and event-share features to predict protest intensity in European datasets, reducing mean squared error by 97% relative to baseline time series models, though the approach failed to generalize effectively to US contexts, illustrating region-specific distributional shifts. Chaves et al. [19] compared text-only GDELT features (for example, keyword TF-IDF vectors, sentiment) against traditional conflict history predictors in the Random Forest and XGBoost classifiers, finding that current news signals nearly matched the ROC-AUC of the models using historical event data. Their LSTM regression further demonstrated that sequential embeddings of daily news could forecast conflict fatality rates with competitive accuracy.

Across these modeling paradigms, several patterns emerge. First, short-term predictive accuracies generally fall within the range 70% to 90%, with sequence models (HMM, LSTM) and graph-based features providing the largest gains over simple count baselines. Second, models often require careful regional calibration: without context-specific retraining, performance degrades markedly in low-media or culturally distinct environments. Finally, all approaches remain vulnerable to noisy inputs—duplicate records, misclassifications, and geolocation errors, underscoring the necessity of rigorous preprocessing and hybrid human-machine validation to ensure robust and generalizable forecasts.

6 Construction of Unrest Indices and Analytical Applications

Researchers have transformed raw GDELT events into a variety of unrest indices and analytical tools, enabling real-time monitoring, econometric modeling, and policy evaluation. Barrett et al. [1] introduced the Real-time Social Unrest Index (RSUI) by normalizing daily protest counts by total media volume and average tone, then applying Bayesian smoothing to highlight statistically significant deviations. RSUI successfully flagged peaks during the 2019 Hong Kong protests and the January 6th, 2021 US Capitol attack, with country-level unrest probabilities climbing by approximately three percentage points at peak.

Machine learning methods have been equally influential. Voukelatou et al. [31] trained a Random Forest model on GDELT event frequencies and tone features to predict Global Peace Index scores. Their model achieved an out-of-bag correlation of 0.67, although residual analysis revealed larger errors in regions with sparse media coverage.

GIS-enabled dashboards like SURGE integrate GDELT with OpenStreetMap infrastructure and terrorism incident layers. Joshi et al. Joshi et al. [11] applied kernel density estimation to fused datasets, creating interactive hotspot maps for South Asia that allow users to filter by event type, date range, and socioeconomic indicators. This real-time spatial analysis has informed targeted relief and policing strategies.

Econometric studies incorporate these indices as dependent variables in causal frameworks. De Cadenas-Santiago et al. [18] used a structural vector autoregression (SVAR), a VAR model augmented with contemporaneous restrictions to identify distinct "repression" and "protest" shocks, in monthly GDELT tone indices and macroeconomic controls, uncovering a U-shaped repression–protest relationship with significant regional spillovers. Iacoella et al. [32] used panel regressions on US county-level lockdown stringency and inequality data, finding that stricter COVID-19 restrictions increased protest probability by 14 percentage points in high-inequality counties. Quaranta's [2] event-GARCH modeling linked European austerity shock announcements to quarter-lagged protest surges from 2000 to 2014, quantifying rapid protest escalation after economic policy changes.

Cross-national burden-shifting studies further extend the reach of GDELT. Sergoyan et al. [33] regressed Azerbaijan's protest and rally counts in global oil price shocks, demonstrating diversionary spikes timed around leadership summits. Baule [34] combined the centrality measure of the event network with demographic covariates in spatial panel models to map US protest clusters, revealing pronounced activity in metropolitan areas of swing states. Gooch et al. [35] computed the Shannon entropy in local language GDELT sources in Africa and Asia, identifying contested narrative environments where the competition for the largest power is most intense.

Finally, methodological refinements are enhancing the index construction itself. Mast et al. [17] quantified "geospatiality" by modeling how different thematic topics (for example, migration and elections) affect the likelihood of extracting usable geotags, informing topic-weighted geoparsing pipelines. Tun et al. [36] analyzed geolocated tweets from 2018 to 2021 Central American migrant caravans using transformer-based sentiment classifiers, then regressed sentiment intensity against GDELT media salience to uncover cross-border opinion dynamics. Murali et al. [37] applied a Bayesian norm-inference algorithm to sequences of bilateral events in GDELT and derived prescriptive rules; for example, they found that mediation actions by an actor are typically followed by cooperative responses. These inferred norms outperformed the baseline discrete-event simulations by a Bayes factor of 12.5.

Together, these indices and applications illustrate how, through careful methodological design and validation, GDELT can be used as a powerful early warning, analytical, and policy evaluation tool in the study of social unrest and political risk.

7 Multi-Source and Advanced Forecasting Frameworks

Multi-source forecasting frameworks represent critical advances in computational social science, combining GDELT's global event stream with diverse external data to overcome its inherent limitations and enhance predictive capabilities. These systems demonstrate that heterogeneous data fusion significantly improves the robustness and applicability of early warning tools for political instability.

The EMBERS project is a leading example of such integration, combining GDELT with Twitter, blogs, Tor activity, web search trends, and economic indicators to forecast protests across Latin America. Korkmaz et al. [21] found that multi-source models achieved F1 scores between 0.68 and 0.95, with social media and news data offering the most predictive power, despite some volatility introduced by the former. Saraf and Ramakrishnan [38] contributed autoGSR to EMBERS, improving the detection of civil unrest events through machine learning automation. The modular pipeline of feature extraction, inference, and alert generation in EMBERS allowed it to successfully forecast major events in Brazil (2013) and Venezuela (2014), illustrating the effectiveness of data fusion and analyst feedback.

Interactive platforms like SURGE offer real-time spatial analysis. Joshi et al. [11] integrated GDELT with terrorism databases, OpenStreetMap infrastructure, and socioeconomic indicators in a dashboard customized for South Asia. SURGE enables hotspot detection, event normalization, and identification of vulnerable districts based on a layered understanding of triggers and facilitators in social unrest.

Advanced neural architectures have further extended predictive modeling. De Oliveira et al. [20]introduced the Graph Language Model (GLM), merging event graphs with transformer-based embeddings to capture semantic and relational features. GLM uses dates, actors, and locations as nodes to model complex event interdependencies, achieving high recall (0.85–0.88) and precision (0.75–0.77) to predict unrest in Hong Kong, USA, France, and Ukraine.

The news and social media sources offer different advantages. Wu and Gerber [23] showed that Twitter excels in next-day predictions, while GDELT captures slightly longer lead signals, such as government actions. Hybrid systems can leverage this complementarity to improve accuracy and timing.

Domain-specific integration has also improved the effectiveness of GDELT. Ndlovu et al. [22] combined GDELT with OSINT sources such as power outages and wage dispute reports to analyze the drivers of unrest in South Africa, identifying 11 key risk factors. Sun et al. [25] incorporated Chinese investment data to map political risk along the Belt and Road Initiative, highlighting areas of regional instability.

Recent innovations explore novel strategies to address GDELT's shortcomings. Macis et al. [39] applied anomaly detection to spot instability through broken temporal trends. Xu and Sun [40] provided a comprehensive overview of event prediction frameworks, underscoring a shift toward multi-modal data fusion and systematized social risk assessment.

These advanced frameworks demonstrate how the integration of GDELT with social, economic, geospatial, and infrastructural data leads to more accurate, context-sensitive early warning systems. However, they also require substantial engineering and methodological rigor to manage the complexity of data integration and ensure operational reliability.

8 Discussion and Future Directions

Our systematic review of fifty GDELT-based forecasting studies reveals substantial methodological progress alongside persistent challenges. Models spanning Hidden Markov frameworks to Graph Language Models now achieve 70% to 90% country-level accuracy, yet performance often varies by context: EMBERS-style multi-source pipelines excel in Latin America, but underperform in low-media regions [10], and transformer and graph-based systems remain sensitive to GDELT's automated coding errors ("∼75%" agreement) [16]. Media bias further distorts input, as English-language and urban outlets dominate, underrepresenting rural and non-Western events and producing weak spatial correlations ($\rho \approx 0.2$–0.3) even when temporal alignment is moderate ($\rho \approx 0.3$–0.6) [6, 7, 26, 28]. Ethical considerations are also underdeveloped where most systems lack differential privacy or bias audits, operating as opaque "black boxex" that risk enabling surveillance and undermining stakeholder trust. Moreover, theoretical work remains descriptive, with only a handful of econometric VAR and event-GARCH studies examining the dynamics of the repression reaction and policy effects [2, 18]. Finally, operationalization beyond prototypes is rare: only SURGE [11] and EMBERS have become live dashboards, highlighting the need for robust, context-sensitive deployments.

To address these gaps, we propose six intertwined priorities. First, **bias correction** must leverage localized sources by incorporating local-language media, NGO reports, and crowd-sourced updates to counteract urban skew in western countries and improve rural event detection [17]. Second, **hierarchical real-time models** should move beyond country-month aggregates to district-level, hourly monitoring via dynamic Bayesian networks or hierarchical LSTMs. Third, **explainable, ethical AI** demands embedding interpretability tools (e.g., attention or SHAP visualizations) and privacy preserving mechanisms (e.g., differential privacy) to ensure transparency and guard against misuse. Fourth, **multi-modal data fusion** can enrich unrest signals by combining GDELT events with satellite imagery, telecom mobility metrics, financial transactions, and climate indicators [40]. Fifth, systems must adopt **adaptive learning** to handle concept drift by implementing continual retraining, drift detection, and automated validation as media ecosystems and protest tactics evolve. Finally, integrating **causal and counterfactual analysis** through structural equation models, synthetic controls, or counterfactual simulations will allow the assessment of policy interventions and guide mitigation strategies.

By integrating these directions within an ethically grounded interdisciplinary framework, the field can transform GDELT from a descriptive event log into a robust, actionable platform for early warning, policy evaluation, and the protection of democratic norms.

Due to the 10-page submission limit, the full bibliography is not included here; it is available upon request.

[pdflatex,sn-mathphys-num]sn-jnl