

Hate Speech as a Conflict-Period Marker: Evidence from Arabic Telegram in Sudan

Tara Pai

Harvard University, USA
tarapai@college.harvard.edu

Poli Nemkova

Columbia University, USA
poli.nemkova@columbia.edu

Abstract

We investigate whether hate-speech intensity on public Telegram channels serves as a conflict signal, using 1,514 days of Arabic-language posts from Sudan aligned with Armed Conflict Location and Event Data Project (ACLED) conflict-event records. Applying a pretrained multilingual hate-speech classifier, we construct a daily intensity index and analyze its relationship to political violence using within-regime analysis and autocorrelation-robust moving-block bootstrap inference. Hate-speech intensity increases approximately four-fold over the months following war onset and remains elevated, but day-to-day fluctuations do not predict short-term changes in violence. Within the post-war period, the index is positively associated with explosions and remote violence ($r = +0.37$) and negatively associated with protests ($r = -0.25$) and riots ($r = -0.30$), with all three associations surviving multiple-comparison correction. These patterns also hold for continuous classifier-score measures, although the regime-level increase is closely intertwined with growth in posting volume and active-channel participation. We further show that pooling across conflict regimes inflates effect sizes, while naive inference on serially dependent time series overstates statistical significance. Hate-speech intensity should therefore be interpreted as a descriptive marker of the conflict-related Telegram environment rather than as an early-warning predictor of imminent violence.

1 Introduction

The relationship between online hate speech and political violence is a central question in computational social science. Research on the Rohingya crisis documented how dehumanizing Facebook rhetoric preceded and accompanied mass violence¹, while broader evidence links social-media

hate speech to real-world hate crimes (Müller and Schwarz, 2020). Yet most prior work examines Western platforms, high-resource languages, and retrospective settings, leaving unclear whether signals from low-resource languages and active conflicts can support operational monitoring.

Telegram² is widely used in contemporary conflicts by militaries, armed groups, diaspora communities, and civilians (Wich et al., 2021; Nemkova et al., 2023; Bawa et al., 2025). Its largely unmoderated public channels and reach in regions with limited traditional media coverage make it a valuable but understudied source of conflict signals.

We ask whether hate-speech intensity on public Telegram channels tracks political violence and, if so, in what form. Using 1,514 days of Arabic-language Telegram posts from Sudan aligned with records from the Armed Conflict Location and Event Data Project (ACLED)³, we construct a daily hate-speech intensity index and assess its predictive and descriptive relationship with violence.

We find that hate-speech intensity serves as a descriptive *conflict-period marker*: it increases approximately fourfold over the months following war onset but does not predict short-term changes in violence. Within the post-war period, it is positively associated with explosions and remote violence and negatively associated with protests and riots, while relationships with other violence types vary across phases. The regime-level increase is robust across classifier-derived measures but closely intertwined with growth in posting volume and active-channel participation. Channel proliferation shows only weak exploratory associations and does not provide evidence of early warning.

These findings inform organizations using social media for conflict monitoring and contribute to

¹A Genocide Incited on Facebook,
With Posts From Myanmar Military:

<https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html>

²<https://telegram.org/>

³<https://acleddata.com/>

research on NLP for atrocity prevention (Benesch, 2012; Olteanu et al., 2018; Karamolegkou et al., 2026).

2 Literature Review

Prior work links hateful online rhetoric to hate crimes, hostility, and political violence (Müller and Schwarz, 2020, 2023). Online hate may not only reflect social tensions but also facilitate mobilization and lower barriers to offline escalation (Müller and Schwarz, 2020). Dehumanizing language can normalize hostility and weaken norms against aggression (Haslam, 2006; Kteily and Bruneau, 2017), while even implicit identity-based rhetoric may precede real-world attacks and serve as an early indicator of conflict escalation (Müller and Schwarz, 2023; Lupu et al., 2023).

Social media is increasingly used to measure public sentiment, monitor political activity, and forecast crises (Arcila Calderón et al., 2024; Bahrami et al., 2018; Yu et al., 2024). Approaches range from topic modeling and event detection to transformer-based extraction of semantic and emotional signals from multilingual corpora (Li et al., 2022). Such data can complement conventional event records by supporting forecasts of protests, civil unrest, and extremist mobilization, particularly when official reporting is delayed or incomplete (Hürriyetoglu et al., 2021). Text-based conflict forecasting also extends beyond social-media platforms. Mueller and Rauh (2018), for example, show that newspaper text can provide information for predicting political violence, illustrating the broader potential of extracting conflict-relevant signals from language. Our analysis examines whether a related signal is observable in a lower-resource, actively contested Telegram environment.

Most conflict-forecasting research focuses on Twitter/X, Facebook, and Reddit, leaving multilingual platforms such as Telegram comparatively understudied (Blas et al., 2025; Scheffler et al., 2021). Telegram’s public channels and limited moderation make it important for conflict-related communication, while its linguistic environment differs from mainstream social media (Rieskamp et al., 2024; Rogers, 2020). We address this gap by applying a multilingual XLM-RoBERTa hate-speech classifier to Arabic Telegram posts from Sudan and testing whether fluctuations in hateful rhetoric precede or accompany independently reported conflict events.

3 Method

3.1 Data

Telegram corpus. We analyze posts from 50 public Telegram channels relevant to Sudanese political and conflict-related discourse from January 2022 to February 2026, retrospectively collected via the official Telegram API (Appendix A1). Channels were identified through keyword searches on public Telegram channel aggregators and manual review of Sudan- and conflict-related content, with a minimum threshold of 1,000 subscribers.⁴ The corpus spans 1,514 consecutive days. The mean number of active channels per day increased from 7.5 in 2022 to 38 in 2025, reflecting variation in channel activity over time and potentially platform growth, conflict-related mobilization, and corpus-composition effects. Posts were language-detected and restricted to Arabic.

Conflict ground truth. We use records from the Armed Conflict Location and Event Data Project (ACLED; Raleigh et al., 2010), a structured dataset of political violence and protest events that provides independently coded information on conflict events, locations, actors, and fatalities. We use Sudan records over the same period and aggregate them to daily event and fatality counts. ACLED distinguishes six subtypes: battles, violence against civilians, explosions/remote violence, protests, riots, and strategic developments.

Hate speech classification. We apply the pre-trained multilingual XLM-RoBERTa hate-speech classifier (Nayak, 2025) to each Telegram post. We selected this model as an openly available, off-the-shelf classifier with Arabic support, allowing us to construct a reproducible hate-speech measure without training on the corpus under study. This keeps the analysis focused on the temporal behavior of an externally defined signal rather than on optimizing a classifier for the present dataset. We do not assume that the model is optimal for Sudanese Arabic; comparison with Arabic-specific hate-speech models remains an important direction for future validation. We define `hate_flag` = 1 when the predicted hate-speech probability exceeds 0.5.

Classifier validity. The classifier was not developed specifically for conflict-related Sudanese Ara-

⁴Data collection was restricted to public channels; no private chats or user-identifiable information were accessed. All data were used solely for research analysis, and no models were trained on the corpus, in compliance with Telegram’s Terms of Service.

	Pred. Hate	Pred. Non-Hate
Actual Hate	6	5
Actual Non-Hate	2	187

Table 1: Confusion matrix for the multilingual XLM-RoBERTa classifier on a manually annotated sample of 200 Telegram posts (5.5% hate speech). Precision = 75.0%, recall = 54.5%, and F1 = 63.2%.

bic, including regional dialects, militia terminology, and war-register vocabulary. Manual evaluation on a random sample of 200 posts yielded 75.0% precision, 54.5% recall, and an F1 score of 63.2% (Table 1). Only 11 posts in the validation sample were manually labeled as hate speech, making the recall estimate uncertain. Moreover, the validation sample was not stratified by conflict regime, so we cannot rule out changes in classifier performance as the linguistic register of the corpus changed after war onset. Misclassification may therefore introduce both measurement noise and systematic bias into estimated hate-speech prevalence.

We consequently interpret the daily index as a classifier-derived proxy for observed rhetorical patterns rather than as a calibrated estimate of hate-speech prevalence. The main descriptive pattern is also recovered using continuous classifier-score summaries, reducing concern that the result is an artifact of the 0.5 decision threshold; however, these robustness checks do not address possible domain- or regime-specific classification error.

3.2 Index Construction

We define daily hate-speech intensity as

$$\text{hate_fraction}(t) = \frac{\text{hate_posts}(t)}{\text{posts}(t)}, \quad (1)$$

which normalizes for daily posting volume. The three days with no observed posts are treated as missing for hate-speech measures. We additionally track the daily mean, median, standard deviation, and 90th percentile of classifier probabilities as threshold-independent robustness measures, and `unique_channels` as a platform-activity indicator. For correlation analyses, series are smoothed using a trailing 7-day window; first differences therefore represent changes in the trailing-window averages.

3.3 Analytical Approach

We use five complementary analyses. **(1) Temporal association:** we correlate hate-speech intensity at time t with ACLED outcomes at $t+k$, for

$k = 0, \dots, 28$, using both levels and first differences. **(2) Regime comparison:** we split the panel at war onset (15 April 2023⁵) and estimate the pre-/post-war mean difference using a moving-block bootstrap (2,000 iterations; primary block length 14 days, with 7-, 21-, and 28-day sensitivity checks). We also compare symmetric 30–365-day windows around onset. **(3) Within-regime specificity:** we estimate correlations with each violence subtype separately within the acute-war, sustained-conflict, and combined post-war periods, applying Bonferroni correction. **(4) Platform-activity sensitivity:** interrupted time-series models with HAC inference estimate post-onset level changes with and without controls for log posting volume and active-channel count; we also examine first-differenced channel-count associations. **(5) Measurement robustness:** we repeat key analyses using continuous classifier-score summaries and after excluding low-volume days.

4 Results

4.1 Conflict Onset Reshapes the Observed Telegram Environment

Table 2 characterizes three phases: pre-war (January 2022–April 2023), acute war (April–December 2023), and sustained conflict (January 2024–February 2026). Mean `hate_fraction` increased from 0.015 pre-war to 0.061 post-onset, while active channels grew from 7.5 to 37.2 per day. The dominant ACLED subtype also shifted from protests pre-war to battles during acute war and explosions/remote violence during sustained conflict.

The aggregate pre/post difference was +0.0459 (95% block-bootstrap CI [0.0420, 0.0512]), corresponding to a $4.02\times$ increase (95% CI [3.46, 5.03]), and was stable across 7–28-day block lengths. However, the increase was gradual rather than an immediate discontinuity: the post/pre ratio was 0.73 within 30-day windows around onset, 1.66 within 90 days, and 3.47 within one year. The pattern also appeared in threshold-independent measures (mean, median, and 90th-percentile classifier probabilities) and remained positive, though attenuated, after excluding low-volume days.

Within-period correlations between `hate_fraction` and daily event counts were not significant under moving-block-bootstrap inference (pre-war: $r = -0.100$, $p_{\text{boot}} = 0.39$;

⁵Beginning of the Sudanese Civil War

post-onset: $r = +0.158$, $p_{\text{boot}} = 0.11$). Thus, the large regime-level difference does not imply dynamic coupling between hate speech and daily violence. Naive Pearson inference produced substantially smaller p -values, illustrating how serial dependence inflates significance in smoothed conflict time series.

4.2 Hate Speech Differentiates Violence Types Within Active Conflict

Because violence composition changes across regimes (Table 2), full-series subtype correlations are inflated by the pre/post-onset shift. We therefore estimate associations within each regime using moving-block-bootstrap inference and Bonferroni correction across six subtypes ($\alpha = 0.0083$).

Within the combined post-onset period, hate-speech intensity was positively associated with explosions/remote violence ($r = +0.373$) and negatively associated with protests ($r = -0.253$) and riots ($r = -0.299$; all $p_{\text{boot}} < 0.0083$). The same signs appear during sustained conflict, although only protests and riots survive correction in that phase. The pattern was also recovered using the daily mean and 90th-percentile classifier probabilities, indicating that it is not specific to the 0.5 threshold.

These effects are moderate and nonstationary. Battles, for example, shift from a negative acute-war association ($r = -0.405$) to a positive sustained-conflict association ($r = +0.541$), but are unrelated in the combined post-onset period. The evidence therefore supports a specific contrast—positive association with explosions/remote violence and negative associations with protests and riots—rather than a uniform relationship with organized violence.

4.3 A Conflict-Period Marker Intertwined with Platform Activity

Level correlations between hate_fraction and conflict events are high and nearly flat across lags ($r = 0.619$ at lag 0 and $r = 0.570$ at lag 28). After first-differencing, however, correlations between changes in the trailing-window averages are near zero at all lags ($r = -0.021$, $p_{\text{boot}} = 0.49$ at lag 0). The index therefore distinguishes conflict regimes but does not predict short-term changes in violence.

Controlling for posting volume, hate fraction retains a modest full-series partial association (partial $r = +0.209$, $p_{\text{boot}} = 0.007$, 95% CI $[+0.06, +0.39]$), but the post-onset partial asso-

ciation is not distinguishable from zero. Interrupted time-series models likewise estimate a post-onset level change without platform controls ($\beta = 0.0415$, $p < 0.001$), but not after controlling for log posting volume ($\beta = 0.0030$, $p = 0.52$), active channels ($\beta = 0.0041$, $p = 0.35$), or both ($\beta = -0.0022$, $p = 0.64$). These results indicate that the regime-level increase in hate-speech intensity is closely intertwined with changes in platform activity and channel composition. However, the aggregate analysis cannot determine whether the increase reflects changes in rhetorical content within persistent channels, changes in the composition of participating channels, or both. Controlling for posting volume and active-channel count does not identify the underlying source of the change. We therefore interpret the post-onset increase as a property of the observed Telegram environment rather than as evidence of an independent within-channel content shift.

4.4 Active-Channel Participation: An Exploratory Platform-Dynamics Observation

Changes in the number of active channels show small positive associations with conflict events ($r = +0.073$, $p_{\text{boot}} = 0.017$ at lag 0; $r = +0.072$, $p_{\text{boot}} = 0.018$ at lag 14). These associations do not survive correction across the tested lags, and variation in active-channel counts may reflect differences in channel activity, availability, deletion, or posting frequency over time. We therefore report the pattern as exploratory rather than as evidence of early warning. Establishing whether changes in active-channel participation carry genuine leading information requires replication across conflicts and more targeted event-study designs around conflict onset.

5 Conclusion

Using 1,514 days of Arabic Telegram posts aligned with ACLED data, we examined whether hate-speech intensity tracks political violence in Sudan. Daily fluctuations did not predict short-term violence after accounting for the shared conflict-period trend. Instead, hate speech functioned as a *conflict-period marker*: it increased gradually after war onset and remained elevated, although this shift was closely intertwined with posting volume and active-channel participation.

Within the post-onset period, hate-speech inten-

	Pre-war Jan'22–Apr'23	Acute war Apr–Dec'23	Sustained Jan'24–Feb'26
Days	469	261	784
Hate fraction	0.015	0.051	0.064
Unique channels/day	7.5	25.2	37.2
Events/day	7.0	18.2	15.5
Fatalities/day	5.3	63.7	49.6
Dominant violence type	Protests	Battles	Explos./Remote

Table 2: Descriptive characteristics of the three conflict regimes. Hate-speech intensity, active-channel counts, and conflict composition differ substantially across periods.

Subtype	Acute war <i>n</i> = 261	Sustained <i>n</i> = 784	Post-war (all) <i>n</i> = 1045
Explosions/Remote	+0.638*	+0.223	+0.373*
Viol. vs. Civilians	−0.100	+0.294*	+0.242
Battles	−0.405	+0.541*	−0.082
Strategic Dev.	—	−0.343*	−0.134
Riots	−0.404	−0.221*	−0.299*
Protests	−0.282	−0.211*	−0.253*

Table 3: Within-regime correlations between hate-speech intensity and ACLED subtypes (trailing 7-day smoothing; moving-block bootstrap, 2,000 iterations). *Significant after Bonferroni correction ($p_{\text{boot}} < 0.0083$). Strategic developments are omitted during acute war because of insufficient non-zero observations. Associations with explosions/remote violence, protests, and riots persist post-onset; other relationships vary across phases.

sity was positively associated with explosions and remote violence and negatively associated with protests and riots, while other relationships varied across phases. These associations are descriptive rather than causal and should not be interpreted as evidence that hate speech causes violence or provides reliable early warning of imminent events.

Importantly, the aggregate data do not establish whether the post-onset increase reflects changes in rhetoric within persistent channels, changes in channel composition, or both. Practitioners should therefore treat elevated scores as a descriptive indicator of the observed conflict-related Telegram environment, not as an independent predictor or a basis for individual- or community-level decisions. Future work should validate the measure on Sudanese Arabic, distinguish within-channel change from channel-composition effects, and evaluate whether comparable signals generalize across conflicts and platforms.

Ethics Statement

This work analyzes publicly accessible Telegram channels and does not include private conversations or personally identifiable information. The study is intended to support aggregate conflict monitoring and humanitarian research rather than the surveillance of individual users, channels, or communities. In particular, the hate-speech index is designed

to characterize changes in the aggregate observed communication environment; it is not intended to identify individuals, assign political affiliation, or determine whether a particular user or community poses a security threat.

Our findings should not be interpreted as evidence that online hate speech predicts imminent violent events or that hate speech causes violence. Instead, the observed relationships describe aggregate patterns within a single conflict setting. An automated hate-speech signal could nevertheless be misused to monitor political opponents, target communities based on their online expression, suppress legitimate political speech, or make high-stakes decisions about individuals or groups. These risks are heightened by the classifier’s limited validation, the possibility of systematic errors across dialects and conflict periods, and the demonstrated dependence of the aggregate signal on channel participation and platform activity. A high hate-speech score should therefore not be treated as evidence that a particular community is dangerous, nor as a basis for enforcement, censorship, targeting, or allocation of coercive resources.

Positive impact. In this setting, positive impact includes not only developing new monitoring capabilities, but also identifying when automated NLP signals should not be trusted for high-stakes

decisions. Demonstrating that classifier-derived hate-speech intensity is a conflict-period marker rather than a reliable early-warning signal can help humanitarian organizations interpret social-media dashboards more cautiously, avoid false alarms and misallocation of monitoring resources, and retain human judgment in operational decision-making.

Accordingly, any future monitoring use should include safeguards such as aggregation over groups and time rather than individual-level scoring, human review before consequential interpretation, explicit documentation of classifier uncertainty and validation limitations, and prohibition of automated decisions based solely on the hate-speech index. The signal should be interpreted alongside independent conflict data and contextual expertise rather than used as a standalone decision-making mechanism.

By clarifying both the value and the limits of this signal, the study supports more responsible deployment of NLP in conflict-monitoring systems.

Limitations

Several limitations bound the interpretation of our findings.

Single-country scope. All results derive from Sudan; generalization across conflicts, languages, and platform ecosystems requires multi-country replication.

Classifier validity. The pretrained multilingual classifier was not developed for conflict-specific Sudanese Arabic and exhibited moderate recall in our manual evaluation (Table 1). Misclassification may therefore introduce measurement noise or systematic bias into the estimated prevalence. Because the validation sample contained few positive examples and was not stratified by conflict regime, we also cannot rule out temporal variation in classifier performance as the linguistic register changed after war onset. We therefore interpret the resulting index as a classifier-derived proxy for observed rhetorical patterns rather than a definitive measure of hate-speech prevalence.

Platform activity and corpus composition. The post-war increase coincides with substantial growth in posting volume and active-channel participation. In interrupted time-series sensitivity analyses, the estimated post-onset level change is no longer distinguishable from zero after controlling for either

measure. Because the aggregate panel cannot separate within-channel rhetorical change from variation in which channels are active and contributing posts at a given time, the index should be interpreted as characterizing the observed Telegram environment rather than as evidence of an independent content effect.

Temporal confounding. Strong level correlations are largely driven by regime-level change and are attenuated after first-differencing. Pooling across periods inflates effect sizes, while serial dependence inflates naive significance estimates. The estimated pre/post difference also decreases when restricting analysis to higher-volume days.

Causal identification. Our analysis is correlational and does not establish whether hate speech causes violence, responds to it, or reflects shared conflict dynamics.

Acknowledgments

This study was supported by the National Science Foundation (NSF) Research Experiences for Undergraduates (REU) Site: Making Generative Artificial Intelligence Responsible (Award #2447834) at the University of North Texas and the University of North Texas AI Summer Research Program.

References

- Carlos Arcila Calderón, Patricia Sánchez Holgado, Jesús Gómez, Marcos Barbosa, Haodong Qi, Alberto Matilla, Pilar Amado, Alejandro Guzmán, Daniel López-Matías, and Tomás Fernández-Villazala. 2024. [From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and lgbt communities](#). *Humanities and Social Sciences Communications*, 11(1).
- Mohsen Bahrami, Yasin Findik, Burcin Bozkaya, and Selim Balcisoy. 2018. [Twitter reveals: Using twitter analytics to predict public protests](#). *arXiv (Cornell University)*.
- Apaar Bawa, Ugur Kursuncu, Dilshod Achilov, Valerie L. Shalin, Nitin Agarwal, and Esra Akbas. 2025. Telegram as a battlefield: Kremlin-related communications during the russia-ukraine conflict. *arXiv preprint arXiv:2501.01884*.
- Susan Benesch. 2012. [Words as weapons](#). *World Policy Journal*, 29(1):7–12.
- Leonardo Blas, Diya Saraf, Tanishq Salkar, Nora Adadurova, Luceri Luca, and Emilio Ferrara. 2025. [Large-scale detection of multilingual coordinated activity on telegram](#). *npj Complexity*, 2(1):33.

- Nick Haslam. 2006. [Dehumanization: An integrative review](#). *Personality and Social Psychology Review*, 10(3):252–264.
- Ali Hürriyetoglu, Hristo Taney, Vanni Zavarella, Jakub Piskorski, Reyhan Yeniterzi, Deniz Yuret, and Aline Villavicencio. 2021. [Challenges and applications of automated extraction of socio-political events from text \(case 2021\): Workshop and shared task report](#). *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, page 1–9.
- Antonia Karamolegkou, Angana Borah, Eunjung Cho, Sagnik Ray Choudhury, Martina Galletti, Pranav Gupta, Oana Ignat, Priyanka Kargupta, Neema Kotonya, Hemank Lamba, et al. 2026. Nlp for social good: A survey and outlook of challenges, opportunities and responsible deployment. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5110–5170.
- Nour S. Kteily and Emile Bruneau. 2017. [Darker demons of our nature: The need to \(re\)focus attention on blatant forms of dehumanization](#). *Current Directions in Psychological Science*, 26(6):487–494.
- Quanzhi Li, Yang Chao, Dong Li, Yao Lu, and Chi Zhang. 2022. [Event detection from social media stream: Methods, datasets and opportunities](#). 2022 *IEEE International Conference on Big Data (Big Data)*, page 3509–3516.
- Yonatan Lupu, Richard Sear, Nicolas Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Beth Goldberg, and Neil F. Johnson. 2023. [Offline events and online hate](#). *PLOS ONE*, 18(1):e0278511.
- Hannes Mueller and Christopher Rauh. 2018. Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, 112(2):358–375.
- Karsten Müller and Carlo Schwarz. 2020. [Fanning the flames of hate: Social media and hate crime](#). *Journal of the European Economic Association*, 19(4):2131–2167.
- Karsten Müller and Carlo Schwarz. 2023. [From hashtag to hate crime: Twitter and antiminority sentiment](#). *American Economic Journal: Applied Economics*, 15(3):270–312.
- Samyak Darshan Nayak. 2025. Multilingual hate speech detection using transformer-based models. <https://huggingface.co/samyak159/xlmroberta-hate-speech-detection>. Hugging Face model card.
- Poli Nemkova, Solomon Ubani, Suleyman Olcay Polat, Nayeon Kim, and Rodney D. Nielsen. 2023. [Detecting human rights violations on social media during russia-ukraine war](#). *arXiv preprint arXiv:2306.05370*.
- Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R. Varshney. 2018. The effect of extremist violence on hateful speech online. *arXiv preprint arXiv:1804.05704*.
- Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset. *Journal of Peace Research*, 47(5):651–660.
- Jonas Rieskamp, Mirbabaie Milad, Marie Langer, and Alexander Kocur. 2024. [From virality to veracity: Examining false information on telegram vs. twitter](#). *Hawaii.edu*.
- Richard Rogers. 2020. [Deplatforming: Following extreme internet celebrities to telegram and alternative social media](#). *European Journal of Communication*, 35(3):213–229.
- Tatjana Scheffler, Veronika Solopova, and Mihaela Popa-Wyatt. 2021. [The telegram chronicles of online harm](#). *Journal of Open Humanities Data*, 7.
- Maximilian Wich, Adrian Gorniak, Tobias Eder, Daniel Bartmann, Burak Enes Çakici, and Georg Groh. 2021. Introducing an abusive language classification framework for telegram to investigate the german hater community. *arXiv preprint arXiv:2109.07346*.
- Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2024. [Hate cannot drive out hate: Forecasting conversation incivility following replies to hate speech](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 18:1740–1752.

A Supplemental Material

A.1 Corpus Statistics

Metric Description	Mean	Std. Dev.	Min	Max
<i>Telegram Platform Content Metrics (Daily Aggregates)</i>				
Total Posts (<i>posts</i>)	629.240	562.887	0.000	4335.000
Hate Speech Flags (<i>hate_posts</i>)	42.062	51.576	0.000	540.000
Hate Speech Ratio (<i>hate_fraction</i>)	0.047	0.032	0.000	0.179
Mean Hate Probability (<i>mean_hate_prob</i>)	0.109	0.049	0.000	0.246
Median Hate Probability (<i>median_hate_prob</i>)	0.042	0.027	0.000	0.156
Std. Dev. of Hate Probability (<i>std_hate_prob</i>)	0.163	0.056	0.000	0.342
Maximum Hate Probability (<i>max_hate_prob</i>)	0.852	0.252	0.000	0.998
90th Percentile Hate Probability (<i>q90_hate_prob</i>)	0.295	0.137	0.000	0.982
Mean Classifier Confidence (<i>mean_confidence</i>)	0.912	0.053	0.000	0.988
<i>Platform Dynamics</i>				
Unique Active Channels (<i>unique_channels</i>)	25.932	13.985	0.000	50.000
<i>Offline Conflict Dynamics (ACLED Daily Counts)</i>				
Total Event Count (<i>event_count</i>)	13.310	8.056	0.000	65.000
Total Fatalities (<i>fatalities</i>)	38.280	66.975	0.000	1024.000
Battles (<i>battles</i>)	4.087	4.390	0.000	28.000
Violence against Civilians (<i>violence_against_civilians</i>)	3.180	3.086	0.000	38.000
Protests (<i>protests</i>)	1.729	3.702	0.000	47.000
Explosions/Remote Violence (<i>explosions_remote</i>)	3.240	3.858	0.000	30.000
Riots (<i>riots</i>)	0.083	0.404	0.000	6.000
Strategic Developments (<i>strategic_developments</i>)	0.990	2.404	0.000	30.000
<i>Geographic, Actor, and Intensity Ratios</i>				
Unique Locations (<i>locations</i>)	12.330	7.380	0.000	63.000
Unique States (<i>states</i>)	5.345	2.423	0.000	18.000
Tracked Factions/Actors (<i>actors</i>)	3.546	1.555	0.000	9.000
Violence Events per Post (<i>violence_per_post</i>)	0.055	0.191	0.000	7.000
Fatalities per Event (<i>fatalities_per_event</i>)	2.441	4.003	0.000	68.267

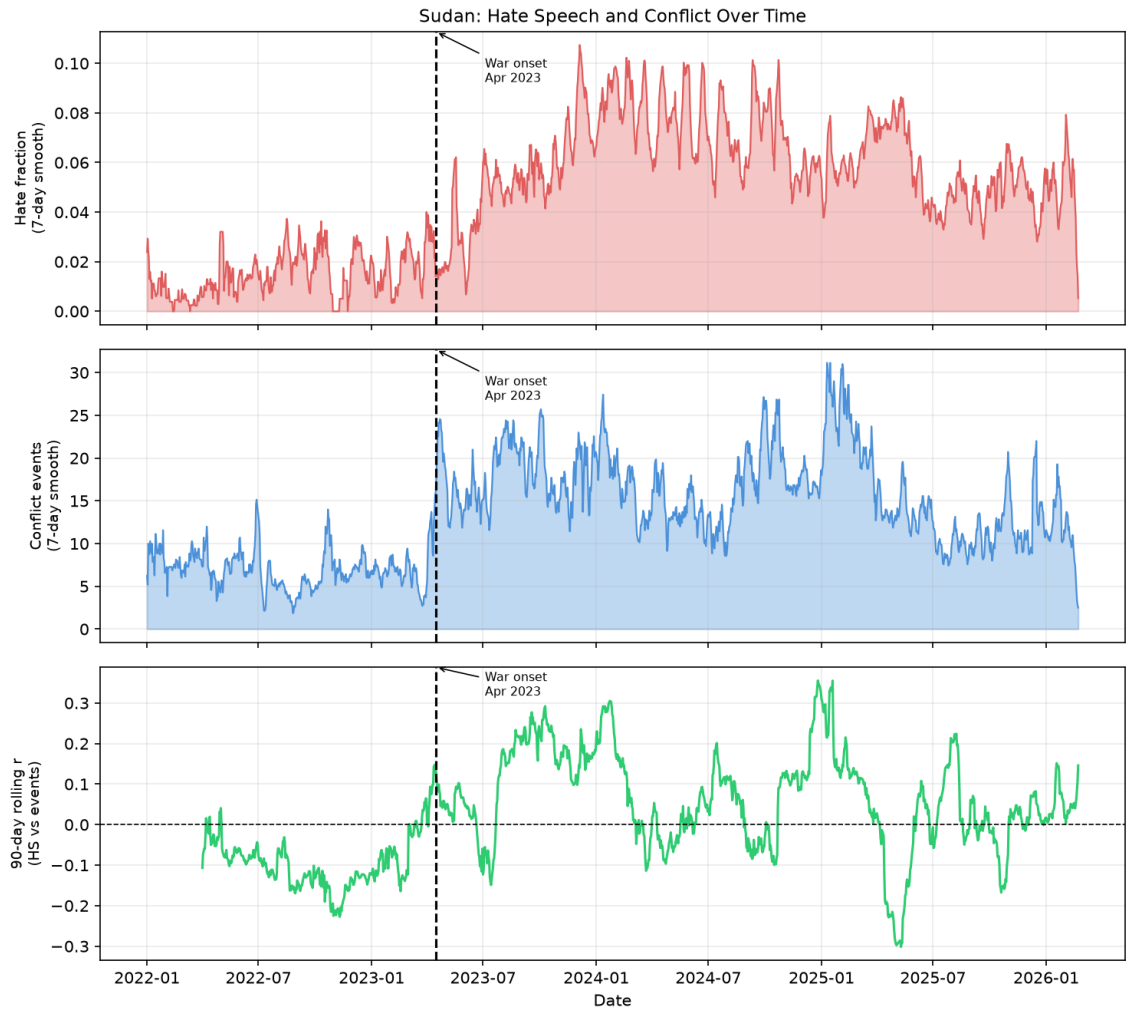
Table A1: Summary statistics of the daily Telegram dataset and aligned ACLED conflict records for Sudan from January 2022 to February 2026 ($N = 1,514$ days). Three dates contained no observed Telegram posts; hate-speech measures for these dates were treated as missing in the analyses. The corpus comprises 50 public Telegram channels; *unique_channels* denotes the number of those channels active on a given day.

A.2 Robustness Analyses

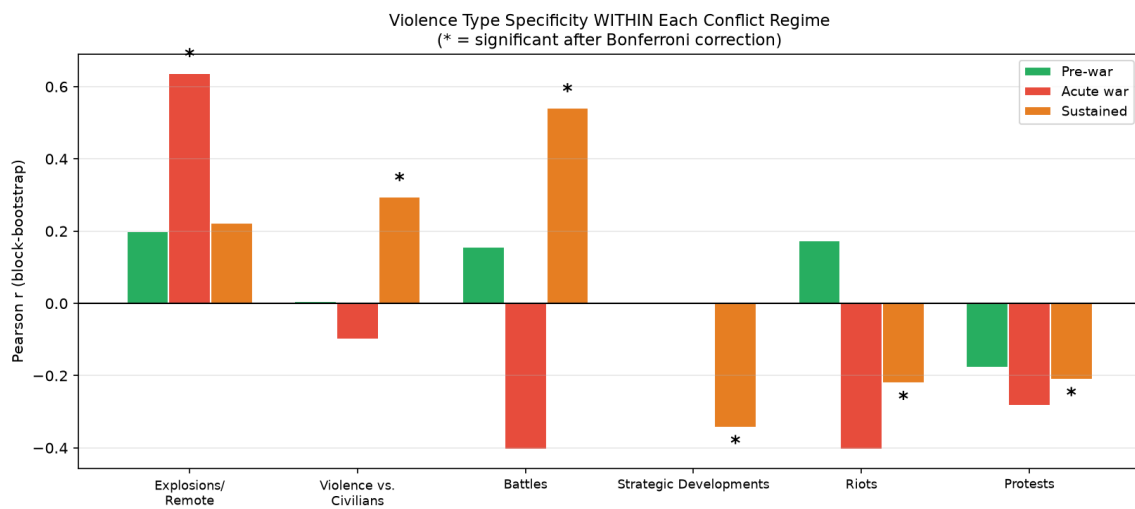
Analysis	Estimate	Interpretation
Hate fraction, pre-/post-onset	0.0152 $\rightarrow$ 0.0611 (4.02 $\times$)	Large descriptive difference between conflict regimes
Block-bootstrap difference	+0.0459; 95% CI [0.0420, 0.0512]	Stable across block lengths of 7, 14, 21, and 28 days
Block-bootstrap ratio	4.02 $\times$; 95% CI [3.46, 5.03]	Confirms a substantial post-onset increase
30-day onset window	Post/pre ratio = 0.73	No immediate increase directly following onset
90-day onset window	Post/pre ratio = 1.66	Increase emerges over the following months
365-day onset window	Post/pre ratio = 3.47	Regime-level difference develops gradually
Mean hate probability	Post/pre ratio = 2.40 $\times$	Threshold-independent robustness
Median hate probability	Post/pre ratio = 3.20 $\times$	Threshold-independent robustness
90th-percentile probability	Post/pre ratio = 2.48 $\times$	Threshold-independent robustness
Minimum 50 posts/day	Post/pre ratio = 3.66 $\times$	Difference remains after removing very low-volume days
Minimum 100 posts/day	Post/pre ratio = 2.39 $\times$	Difference remains positive but is attenuated
ITS without platform controls	$\beta = 0.0415, p < 0.001$	Significant estimated post-onset level difference
ITS controlling for log posts	$\beta = 0.0030, p = 0.52$	No independent post-onset level difference
ITS controlling for active channels	$\beta = 0.0041, p = 0.35$	No independent post-onset level difference
ITS controlling for both	$\beta = -0.0022, p = 0.64$	Regime shift is intertwined with platform activity

Table A2: Robustness analyses of the pre-/post-onset difference in classifier-derived hate-speech intensity. ITS denotes interrupted time-series analysis with heteroskedasticity- and autocorrelation-consistent inference.

A.3 Supplemental Figures



(a) Evolution of hate-speech intensity, conflict events, and their trailing-window changes around war onset.



(b) Within-regime associations between hate-speech intensity and ACLED violence subtypes.

Figure A1: Temporal and type-specific relationships between classifier-derived hate-speech intensity and conflict dynamics. The vertical line in panel (a) marks the onset of open RSF–SAF hostilities on 15 April 2023 and does not imply an instantaneous discontinuity. Stars in panel (b) indicate significance after Bonferroni correction.

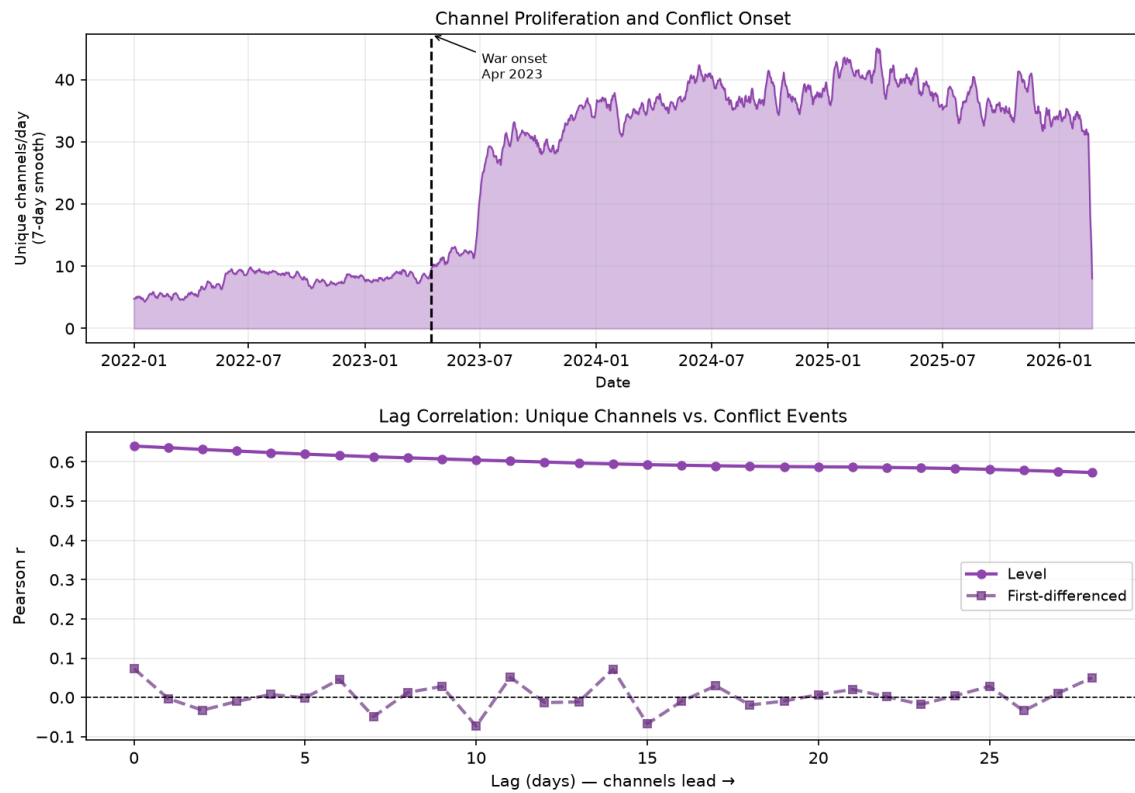


Figure A2: Active-channel participation across the study period and lagged associations with conflict events. First-differenced associations are small and do not survive correction across the tested lags; the result is therefore interpreted as exploratory.

A.4 Examples of Telegram Posts

Paraphrased content	Interpretive gloss	Hate probability
Dehumanizing insult targeting a perceived collaborator	Speaker calls another person a filthy dog and accuses them of collaboration.	0.971
Severe direct personal insult	Speaker addresses another person with a strong insult and emphasizes that both parties understand the insult being made.	0.814
Call for surrender at military sites	Post urges fighters located at several camps and weapons facilities to surrender and turn themselves in.	0.501
Request for reassurance about loved ones	Post asks for updates and reassurance about family members, loved ones, and the army.	0.500

Table A4: Illustrative posts from the Telegram corpus. Text has been paraphrased and identifying details generalized to reduce traceability to specific public channels. Hate probabilities correspond to outputs from the XLM-R hate speech classifier and should not be interpreted as human annotations.