From Diplomacy Discourse to On-the-Ground Events

Project Narrative (Methods and Results Summary)

Berk Sankır

2026

This report consist of 7 main Titles:

1. Motivation and Research Question

This project is motivated by an early-warning question: whether changes in how one country talks about another country are informative about near-future events on the ground. In the Russia–Ukraine war context, diplomatic statements often signal support, condemnation, escalation cues, and framing choices. If these signals move systematically before spikes in protests or military/tension events, they could help anticipate risk in the following weeks.

Research question: Do weekly changes in diplomatic discourse from country i toward country j help predict protest or military/tension events in country j over the next week (and, conceptually, the next 1–4 weeks)?

2. Data Sources

Three public data sources are used:

• GlobalDiplomacyNet (GDN): diplomatic texts and relations across countries and years.

• GDELT Events v2: event records used to derive weekly outcomes (protest and tension/violence families).

• Lowy Global Diplomacy Index: yearly context measures (used as optional context; not required for core results).

All processed data used in the project is shared via the project Google Drive folder (data_processed).

3. Data Processing and Feature Construction

The unit of analysis is a dyad-week panel covering 2019–2024. Each row corresponds to one directed dyad (i→j) in a given week. The core dyads are US–RU, US–UA, and RU–UA. The panel integrates discourse-derived features with event outcomes.

3.1 Diplomatic discourse features

From weekly diplomatic text collections, I compute lexicon-based ratios and scores that capture broad themes and tone. The main feature set includes: war_ratio, peace_ratio, security_frame_ratio, economy_frame_ratio, humanrights_frame_ratio, support_ratio, condemn_ratio, and a tone_support_score. I also track num_docs (the number of documents contributing to the week's discourse signal) to represent volume and to reduce instability in weeks with very low text coverage.

To support next-week prediction, lagged versions of the discourse features (lag1) are created by shifting each dyad's feature values by one week.

3.2 Outcome construction (GDELT)

Weekly event outcomes are derived from GDELT by aggregating event counts by country-week. The primary outcomes are protest_next_week and military_next_week, defined as next-week event counts. For classification analyses, binary outcomes are constructed as 1[count>0].

3.3 Train/validation/test split

To avoid temporal leakage, all modeling uses time-based splits: Train (years ≤ 2022), Validation (year = 2023), and Test (year = 2024). Threshold selection and model decisions are made using Validation, while Test is used only for final reporting.

4. Hypothesis Testing

Before ML modeling, I run hypothesis tests to check whether discourse features and event outcomes show structured relationships. These tests provide interpretability and sanity checks, and they help motivate the predictive tasks.

4.1 H1 — Period difference (Pre-war vs War)

I compare Pre-war (2019–2021) vs War (2022–2024) periods for both binary event rates and event counts. For binary outcomes I use a chi-square test of independence; for count outcomes I use a Mann–Whitney U test.

• Protest (binary): event rate increases from 0.158 to 0.274 (chi-square p = 1.61e-05, odds ratio ≈ 2.02).

• Military/tension (binary): rates are similar (0.326 vs 0.331); chi-square p = 0.869.

• Counts: both outcomes show significant distribution shifts (Mann–Whitney p = 1.61e-03 for military/tension; p = 7.64e-07 for protest).

4.2 H2 — Discourse → next-week military/tension events

To test whether discourse features are associated with next-week military/tension event counts, I estimate one-feature-at-a-time Negative Binomial (NB) GLMs with robust (HC1) standard errors. Models include controls for num_docs and fixed effects for dyad and year, plus a week control. Features are scaled by 10,000 for coefficient stability.

The strongest robust association is for peace_ratio, which is negatively associated with next-week military/tension counts (FDR-adjusted p ≈ 3.26e-06, IRR ≈ 0.987, 95% CI [0.982, 0.992]).

4.3 H3 — Discourse → next-week protest events

I run the same NB-GLM testing approach for protest counts. Results suggest that more supportive discourse is associated with fewer next-week protest events in this dataset, although significance is marginal after multiple-testing correction.

Support_ratio shows the clearest negative relationship (FDR-adjusted p ≈ 0.053, IRR ≈ 0.682, 95% CI [0.518, 0.899]).

5. Machine Learning Modeling

After hypothesis testing, I frame the task as short-horizon early warning using next-week labels. I evaluate both classification (any event next week) and regression/count prediction (how many events next week). Model selection decisions are based on validation performance and then summarized on the held-out 2024 test set.

5.1 Binary prediction (classification)

For binary tasks, I evaluate Logistic Regression (with scaling and class_weight=balanced) and HistGradientBoostingClassifier as baseline models. Feature sets include discourse ratios/scores, optional num_docs, and optional lag1 features. Given class imbalance, I focus on PR-AUC and F1 in addition to ROC-AUC.

5.2 Threshold selection (validation-tuned)

Predicted probabilities are converted to binary decisions using thresholds selected on the 2023 validation set. I report two operating points: (i) VAL_F1_OPT (threshold maximizing F1) and (ii) VAL_PREC≥0.70_MAXREC (a higher-precision constraint, choosing the threshold with the best recall subject to precision ≥ 0.70). Test metrics are reported using these validation-selected thresholds.

5.3 Ablation (feature robustness)

I run an ablation study to isolate the impact of including num_docs and lag1. The final configuration uses lag1 for both tasks. For protest prediction, including num_docs improves results; for military prediction, excluding num_docs performs better.

5.4 Count prediction

For count outcomes, I compare a Negative Binomial GLM baseline against a boosting regression baseline on a log1p-transformed target. The Negative Binomial GLM provides the best overall error (MAE/RMSE) on the test set in this dataset.

5.5 Calibration experiment

As an additional analysis, I test probability calibration using cross-validated calibration on the train+validation set and evaluation on the test set. Isotonic calibration performed poorly in this setting. Sigmoid (Platt) calibration improved Brier score for the military task but did not help protest, and it slightly reduced PR-AUC. Therefore, uncalibrated probabilities are retained as the default for ranking and thresholding, with calibrated probabilities considered optional when well-calibrated risk estimates are required.

5.6 Summary of final ML results (Test 2024)

Binary tasks (VAL_F1_OPT thresholds) on the 2024 test set:

Task	Model	Thr	ROC-AUC	PR-AUC	Precision	Recall	F1
PROTEST_BINARY	HGB-Cls	0.6	0.875	0.698	0.538	0.737	0.622
MILITARY_BINARY	LogReg	0.55	0.894	0.773	0.598	0.961	0.737

Count tasks (Negative Binomial baseline) on the 2024 test set:

Task	Model	MAE	RMSE
PROTEST_COUNT	NegBin(GLM)	0.572	0.934
MILITARY_COUNT	NegBin(GLM)	4.031	5.572

Calibration summary (Test 2024):

Task	Model	Calibration	PR-AUC	Brier
PROTEST_BINARY	HGB-Cls	none	0.704	0.125
PROTEST_BINARY	HGB-Cls	sigmoid(cv=5)	0.685	0.145
MILITARY_BINARY	LogReg	none	0.782	0.173
MILITARY_BINARY	LogReg	sigmoid(cv=5)	0.781	0.139

6. Limitations and Next Steps

This study focuses on a small number of dyads in a specific geopolitical context, so results may not generalize to other regions or to different time periods. Features are lexicon-based and intentionally simple; richer NLP representations (e.g., embeddings or supervised stance classifiers) could improve signal quality. Event labels are imbalanced and depend on GDELT coverage; additional robustness checks could include alternative event taxonomies, different forecast horizons (t+2 to t+4), and expanded dyad coverage.

7. Conclusion

The project suggests that weekly diplomatic discourse signals have measurable early-warning value for next-week on-the-ground events, but the strength of this value depends on the event type. For **military/tension events**, discourse content features (especially condemnation and security/war framing and their short lags) provide a meaningful predictive signal for whether an event will occur next week, indicating that changes in diplomatic tone and framing can serve as a practical risk indicator. For **protest events**, next-week prediction is feasible, but performance relies heavily on **text volume/coverage (num_docs)**, implying that attention intensity is a major driver and that "pure content" effects are weaker. Across both tasks, the models perform better at predicting **event occurrence (binary)** than **event magnitude (counts)**, where a Negative Binomial baseline remained the most stable option and more complex approaches did not improve errors. Overall, the findings support the idea that dyad-week discourse features can contribute to short-horizon early warning—most clearly for military/tension events—while also highlighting important limitations for protest prediction and for count forecasting.