

# Migration Surge Prediction: An Early Warning System

|                                          |                                 |                               |
|------------------------------------------|---------------------------------|-------------------------------|
| <b>Syed Ibrahim Omer</b><br>58776310     | <b>Haseeb</b><br>58791350       | <b>Asadohoo</b><br>58452930   |
| <b>Sameer Ahmed SIDDIQUI</b><br>58771398 | <b>Ayaan Yousuf</b><br>58970130 | <b>Pak Hei Ng</b><br>58650183 |

SDSC 2005 — Introduction to Computational Social Science  
Group 10 Final Report  
April 2026

## Abstract

Governments and humanitarian organisations are routinely caught off-guard by sudden spikes in migration, yet the underlying signals—shifting news coverage, search behaviour, and macroeconomic volatility—are observable months in advance. This study presents a multi-modal early warning system that synthesises over 170,000 news articles, Google Trends search-intent data, and Real Effective Exchange Rate (REER) indicators across 15 high-volume origin countries over a nine-year observation window (2017–2026) to predict surges in both legal immigration (visa issuances) and unauthorized border encounters at one-to-six-month horizons. We evaluate four architectures—a GPU-accelerated Random Forest baseline, a recurrent LSTM with a novel dual-objective *SurgeJointLoss*, a Transformer encoder, and a Horizon-Aware Ensemble that dynamically re-weights models by forecast lead time. Evaluated strictly out-of-time on walk-forward splits (training  $\leq 2022$ , testing 2023+), the ensemble achieves an average F1 score of 0.915 with recall of 0.943, while maintaining recall above 92% at every lead—meaning fewer than one in ten genuine crises goes undetected. These results demonstrate that fusing heterogeneous digital trace data with purpose-built deep learning can transform migration forecasting from a reactive exercise into a proactive planning tool.

## **Contents**

## 1 Introduction

### 1.1 Study Context and Motivation

Global migration is driven by a deeply intertwined web of push and pull factors—geopolitical instability, economic opportunity, conflict, climate events, and family reunification—making the timing and magnitude of migration flows inherently difficult to anticipate. For the roughly 280 million international migrants worldwide (?), the decision to move is rarely sudden; it is incubated by months of deteriorating conditions at home and perceived opportunity abroad. The signals of that incubation period—a spike in “US asylum” searches, a currency devaluation, a surge in conflict-related news coverage—are increasingly observable in real time through digital trace data.

Yet current migration management remains overwhelmingly *reactive*. Governments confront surges only *after* they materialise—scrambling to allocate housing, legal processing capacity, and humanitarian aid when border crossings have already overwhelmed existing infrastructure. In May 2023, U.S. Southwest border encounters peaked at roughly 880,000 in a single month (?)—approximately 18 times the 2018–2019 baseline. Traditional forecasting suffers from what we term a “**precision blindspot**”: models trained on slow-moving historical indicators adequately predict the normal baseline but systematically miss the extreme spikes that actually trigger crises.

This project proposes an **Early Warning System** that addresses this blindspot by synthesising three families of real-time digital signals—(1) news event coverage reflecting geopolitical conditions in origin countries, (2) Google Trends search-intent data capturing migration-related interest, and (3) macroeconomic exchange-rate volatility—into actionable, multi-horizon surge forecasts spanning one to six months ahead.

### 1.2 Research Questions

- **RQ1:** How can multi-modal digital trace data (news coverage, search intent, and macroeconomic indicators) be synthesized to predict sudden surges in migration?
- **RQ2:** How does the predictive accuracy of various architectures (Random Forest, LSTM, Transformer) change across forecasting horizons (1–6 months)?
- **RQ3:** Which data modality—news coverage, Google Trends, or exchange rates—serves as the strongest early indicator of a migration spike?

## 2 Methods

### 2.1 Data Collection and Multi-Modal Ingestion

Our system ingests four data streams aligned to a monthly cadence, covering the 15 highest-volume origin countries for U.S. migration (Afghanistan, China, Colombia, Cuba, Dominican Republic, El Salvador, Guatemala, Haiti, Honduras, India, Mexico, Pakistan, Philippines, Venezuela, Vietnam).

**Ground Truth—Visa Issuances.** Monthly immigrant visa issuance statistics were collected from the U.S. Department of State (?), spanning January 2017 through September 2025. The 108 monthly PDF reports were parsed programmatically using PyMuPDF table extraction; 118 distinct visa classification codes were mapped into seven interpretable categories (family-sponsored, employment-based, special immigrant, humanitarian, returning resident, diversity lottery, and non-immigrant), yielding ~100,000 records. Exploratory analysis confirmed the distribution is Zipfian: the top 20% of countries account for 88.3% of all visas issued, and the top 20% of countries account for 81.7% of all surge months—an empirical regularity exploited to anchor the country scope (see Section 2.3).

**Ground Truth—Border Encounters.** Southwest U.S.–Mexico land border encounter data was obtained from U.S. Customs and Border Protection (?), comprising eight CSV datasets (FY

2019–2026 YTD) merged via Polars (?) lazy evaluation.

**News Articles.** A large-scale automated pipeline collected articles via the Google News RSS API (?) across all 15 countries and 8 migration-relevant topics (e.g., “US visa,” “US asylum,” “CBP One”), producing 120 unique query combinations. Articles were fetched asynchronously with bounded concurrency (16 connections), decoded in batches of 20, and extracted via Trafilatura (?) parallelized across CPU cores. This yielded **170,784** raw articles, of which **104,333** passed quality filters—a 61.1% success rate. Details of the pipeline architecture appear in Appendix ??.

**Google Trends.** Monthly Google Trends time series were collected for eight migration-intent keywords (us\_visa, us\_asylum, green\_card, cbp\_one, usaid, united\_states\_passport, us\_jobs, study\_us) across all 15 countries (?).

**Exchange Rates.** Real Effective Exchange Rate (REER) data was obtained from the IMF (?) for the 6 of 15 countries where data was available (China, Colombia, Dominican Republic, Mexico, Pakistan, Philippines)—a known coverage limitation.

## 2.2 NLP Processing and Feature Engineering

**Embedding Generation.** Each of the 104,333 articles was tokenized (max 8,192 tokens) and encoded into a 768-dimensional embedding using a Jina v5 model (?) accelerated via TensorRT INT4 quantisation (?). TensorRT compilation applied four optimisations: (i) *operator fusion*—combining adjacent GEMM, bias, and activation operations into a single CUDA kernel; (ii) *layer fusion*—merging consecutive matrix-multiply and non-linear layers; (iii) *CUDA graph capture*—recording the full inference computation graph once to eliminate per-step kernel launch overhead; and (iv) *INT4 weight quantisation* for the embedding model and INT8 (W8A16) for the Flan-T5-Large labelling model. These optimisations yielded throughputs of **20,000+ tokens/second** for Jina v5 (sequence length 8,192) and **6,000+ tokens/second** for Flan-T5-Large (sequence length 512), enabling overnight batch processing of the full article corpus. Quantization validation confirmed mean cosine similarity of 0.9999 against the FP32 baseline.

**Event Clustering.** To discover thematic event groups, we applied HDBSCAN (?) on the embeddings via NVIDIA cuML (?) (min\_cluster\_size=10, min\_samples=5, Euclidean metric). Clusters were visualized with UMAP (?) and *t*-SNE (?).

**Autonomous Cluster Labelling.** A TensorRT-optimized Flan-T5-Large model (INT8 W8A16 quantisation) (??) generated 2–3 word descriptive labels for each cluster from five sampled article headlines. Approximately 75% of clusters received meaningful labels (e.g., “Honduras Evangelical Pastors”, “Cuba—Trump Policy”); the remainder fell back to a generic “Migration event update” tag when the output failed constraint checks, and these were retained as-is.

**Headline-Level Scoring.** A lexicon-based scoring function was applied to each article **headline** using a domain-specific vocabulary of 19 positive terms (e.g., “growth,” “peace,” “jobs”) and 27 negative terms (e.g., “crisis,” “conflict,” “violence”):

$$s = \text{clip}\left(\frac{n_+ - n_-}{\sqrt{n}}, -1, 1\right), \quad (1)$$

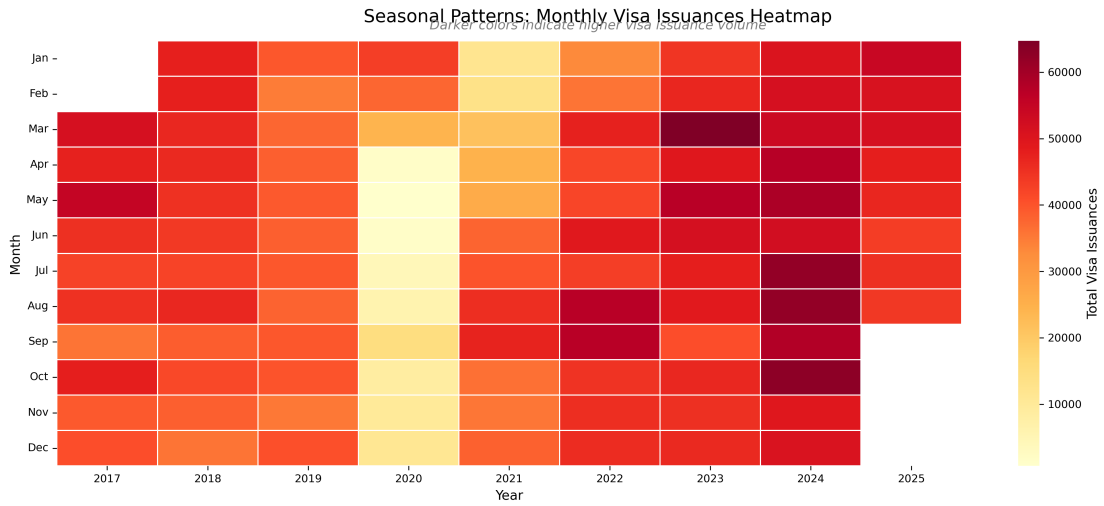
where  $n_+$  is the count of matched positive lexicon terms in the headline,  $n_-$  is the count of matched negative terms,  $n$  is the total token count of the headline, and  $\text{clip}(\cdot, -1, 1)$  clamps the raw score to avoid division-by-small- $n$  inflation. Monthly country–cluster sentiment means were computed by aggregating per-article scores; score shocks were flagged via  $z$ -score thresholds ( $z \leq -1.0$  negative,  $z \geq +1.0$  positive).

**Panel Construction.** All streams were merged at the monthly  $\times$  country level. For each observation we constructed **lag features** ( $t-1$  through  $t-6$ ) for three modalities—visa volume, exchange rate, and news event count—yielding 18 input features, plus **lead targets** ( $t+1$  through  $t+6$ ) of visa volume.

### 2.3 Exploratory Data Analysis

Exploratory analysis of the visa issuance data revealed two structural features that shaped subsequent modelling choices.

**Seasonal Patterns and Pandemic Disruption.** Figure ?? shows the monthly visa issuance heatmap from 2017 to 2025. Volume was broadly consistent across 2017–2019, collapsed during 2020 owing to pandemic-era consulate closures, and rebounded sharply from 2021 onward. The summer and early-autumn months (July–October) of 2023–2024 represent the peak issuance period across the observation window, confirming a seasonal cycle that any predictive model must account for.

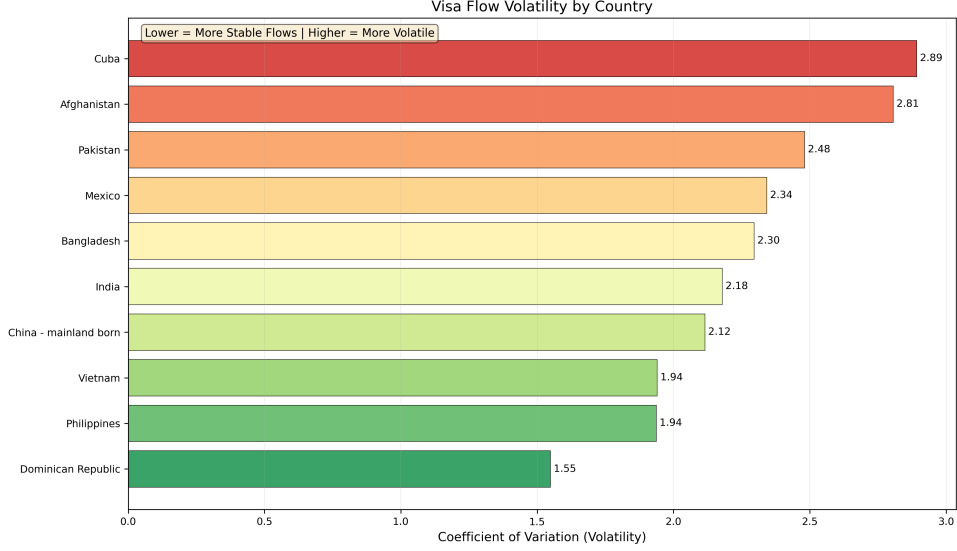


**Figure 1:** Seasonal Patterns: Monthly Visa Issuances Heatmap (2017–2025). Darker red shading indicates higher monthly volume. The light-coloured column at 2020 reflects near-complete consulate closures during the COVID-19 pandemic.

**Zipfian Concentration and Country Volatility.** As noted in Section ??, the cross-country distribution of visa issuances follows a Zipfian law—the top 20% of sending countries contribute 88.3% of all issuances, and 81.7% of all surge months—motivating the focus on 15 high-volume origins. Country-level volatility, measured by the coefficient of variation (CV) of monthly issuances, further identifies which countries are most likely to generate abrupt surges. As Figure ?? shows, Cuba (CV 2.89), Afghanistan (CV 2.81), and Pakistan (CV 2.48) are the three most volatile origins, while the Dominican Republic (CV 1.55) is comparatively stable despite producing the strongest leading signal in the event–correlation analysis.

### 2.4 Lead–Lag Signal Analysis

For each country–signal–lag combination (lags 0–6), Pearson correlation was computed between the signal and visa volume, with  $p$ -values corrected via Benjamini–Hochberg FDR (?) at  $q \leq 0.10$ . Surge months were defined by a dual criterion: volume above the 75th percentile *or* month-over-month growth exceeding 30%. Signal quality was quantified via surge precision and surge lift.



**Figure 2:** Visa flow volatility by country, measured as the coefficient of variation of monthly visa issuances. Higher values indicate more erratic, surge-prone flows. Colour encodes volatility level (red = high, green = moderate).

## 2.5 Predictive Models

**cuML Random Forest.** A GPU-accelerated Random Forest (??) with 50 estimators and max depth 6 served as a non-temporal baseline; six independent models were fit, one per horizon.

**LSTM with SurgeJointLoss.** A two-layer LSTM (?) (hidden dim 64) with learned country embeddings (dim 8) was trained with a novel dual-objective loss function, *SurgeJointLoss*:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{Huber}}(\hat{y}, y) + (1 - \alpha) \mathcal{L}_{\text{BCE}}(\hat{y} - \tau, \mathbf{1}[y > \mu + 1.5\sigma]), \quad (2)$$

where  $\hat{y}$  is the predicted monthly visa volume,  $y$  is the observed volume,  $\alpha = 0.6$  is a fixed mixing weight that places 60% emphasis on volume regression and 40% on surge detection,  $\tau$  is a learned discrimination threshold,  $\mu$  is the rolling mean of historical visa volume,  $\sigma$  is its rolling standard deviation (so  $\mu + 1.5\sigma$  defines the surge boundary),  $\mathcal{L}_{\text{Huber}}$  is the Huber regression loss (?) (robust to extreme outliers), and  $\mathcal{L}_{\text{BCE}}$  is Binary Cross-Entropy applied to the surge indicator. Both neural architectures were implemented in PyTorch (?) and optimised with Adam (?).

**Transformer Encoder.** A two-layer Transformer (??) with 4 attention heads ( $d_{\text{model}} = 64$ ) and learnable positional encodings was trained with MSE loss as a pure attention-based comparator.

**Horizon-Aware Ensemble.** The final model blends predictions from all three architectures with horizon-specific weights that shift trust from RF/LSTM dominance at short leads toward Transformer dominance at long leads (see Appendix ?? for the full schedule).

## 2.6 Validation Strategy

All models were evaluated under a strict **walk-forward, out-of-time (OOT)** protocol: training on data through December 2022, evaluation on 2023+. A month-country observation was classified as a surge if volume exceeded  $1.5\sigma$  above the rolling mean ( $\sim 103$ – $114$  surge instances in the test set). We report Precision, Recall, F1, and RMSE.

### 3 Results

#### 3.1 Multi-Signal Analysis (RQ3)

**News Event Signals.** Of 392 country–event–lag combinations tested, 58 (14.8%) achieved significance ( $q \leq 0.10$ ). The strongest single signal was the Dominican Republic, where a specific cluster showed  $r = 0.617$  at a 3-month lag ( $q = 3.3 \times 10^{-9}$ ). Cuba’s “Trump Policy” cluster showed  $r = -0.595$  at lag 0 ( $q = 3.4 \times 10^{-9}$ ), reflecting immediate U.S. policy impacts. Eleven of 15 countries exhibited at least one significant event signal. Full results appear in Appendix ??.

**Exchange Rate Signals.** Among 6 countries with REER data, 23 of 42 lag rows (54.8%) were significant. The Dominican Republic dominated ( $r = 0.498$ , lag 2,  $q = 2.6 \times 10^{-6}$ ). Overall surge lift was modest (mean 1.009), confirming exchange rates as a complementary rather than standalone signal.

**Google Trends.** Correlations were significant for 66 of 120 rows (55.0%), but a trends-enhanced VAR (??) model *underperformed* the autoregressive baseline (−5.97% mean RMSE for visas), indicating that translating keyword correlations into out-of-sample gains requires non-linear modelling.

**Modality Ranking.** News event coverage emerged as the strongest individual predictor (highest single-signal correlations, most country-specific leads); exchange rates added meaningful signal where available (6 countries); Google Trends showed the broadest coverage but the weakest out-of-sample translation—motivating the multi-modal fusion approach.

#### 3.2 Model Performance

Table ?? presents the aggregate surge-classification performance across all six horizons.

**Table 1:** Average surge-classification performance (Leads 1–6, OOT test set).

| Algorithm                     | Precision    | Recall       | F1 Score     |
|-------------------------------|--------------|--------------|--------------|
| cuML Random Forest            | 0.870        | 0.943        | 0.907        |
| LSTM (SurgeJointLoss)         | 0.895        | 0.820        | 0.855        |
| Transformer Encoder           | 0.892        | 0.917        | 0.905        |
| <b>Horizon-Aware Ensemble</b> | <b>0.887</b> | <b>0.943</b> | <b>0.915</b> |

The ensemble achieves the highest average F1 (0.915) by combining the RF’s high recall (0.943) with the Transformer’s balanced precision–recall profile. The LSTM’s conservative SurgeJointLoss objective yields the highest precision (0.895) at the cost of lower recall (0.820). Per-horizon breakdowns are provided in Appendix ??.

Table ?? reports volume prediction accuracy (RMSE) for the three base models.

The LSTM dominates near-term volume prediction (Leads 1–3) by 10–20 RMSE points; at longer horizons the RF maintains competitive parity.

#### 3.3 Architecture Insights (RQ2)

1. **Random Forest:** highest single-model F1 at Lead 1 (0.97) and recall of 0.943 on average, but precision degrades at longer horizons.
2. **LSTM:** highest precision (0.895 average) thanks to SurgeJointLoss—when it fires an alarm, it is the most likely to be genuine. The trade-off is lower recall (0.820).
3. **Transformer:** superior long-horizon recall (90% at Lead 6) and the highest individual-model F1 at Lead 6 (0.87), capturing slow geopolitical dynamics via self-attention.

**Table 2:** Out-of-time RMSE by model and horizon (lower is better). Bold marks best per row.

| Horizon | Random Forest | Transformer | LSTM          |
|---------|---------------|-------------|---------------|
| Lead 1  | 298.79        | 303.43      | <b>277.40</b> |
| Lead 2  | 341.79        | 343.61      | <b>320.77</b> |
| Lead 3  | 351.55        | 359.20      | <b>346.03</b> |
| Lead 4  | <b>364.01</b> | 381.20      | 381.00        |
| Lead 5  | 428.35        | 407.67      | <b>404.23</b> |
| Lead 6  | <b>401.38</b> | 426.48      | 424.59        |

4. **Ensemble:** by shifting trust from RF/LSTM at short leads toward the Transformer at long leads, it achieves the best F1 at every horizon and maintains **recall**  $\geq 0.92$  across all six months.

## 4 Conclusion

This study has demonstrated that multi-modal digital trace data—when processed through architectures designed to capture surge dynamics—can reliably anticipate migration crises months before they materialise. Across all three research questions the evidence converges on a clear answer: fusing news event coverage, search-intent signals, and macroeconomic indicators into a Horizon-Aware Ensemble yields consistent gains over every single-modality or single-architecture baseline, with the ensemble achieving an average F1 of 0.915 and recall  $\geq 0.92$  at every forecast horizon from one to six months. These results indicate that the “precision blindspot” afflicting traditional reactive migration management is not an inherent feature of the forecasting problem, but an artefact of relying on slow-moving aggregate statistics rather than the rich signal landscape now observable through digital sources. Pre-emptive planning—staffing adjustments, logistics pre-positioning, diplomatic engagement—becomes tractable when high-recall early warnings arrive months in advance.

### 4.1 Practical Implications

- **Lead 1 (F1 0.96):** Immigration processing centres can adjust staffing for the coming month with high confidence.
- **Lead 3 (F1 0.93):** Humanitarian organisations can pre-position supplies and shelter capacity in anticipated surge corridors.
- **Lead 6 (F1 0.86):** National-level budgetary and diplomatic planning can be initiated with sufficient institutional lead time.

The finding that news event signals around U.S. policy announcements serve as the strongest leading indicators suggests that **monitoring origin-country media coverage could function as a low-cost early warning input** even without sophisticated modelling infrastructure.

### 4.2 Limitations

**The Digital Divide.** Google Trends and digital news reflect populations with internet access. The most vulnerable migrants—those displaced by conflict or extreme poverty in regions with limited digital infrastructure—may be systematically underrepresented, creating a blind spot for the most acutely at-risk groups.

**Correlation, Not Causation.** This is a *predictive* model, not a *causal* one. Lead-lag correlations identify temporal associations, but establishing causal mechanisms would require quasi-experimental or instrumental-variable approaches outside the present scope.

**Exchange Rate Data Gap.** REER data was available for only 6 of 15 target countries (?), excluding high-priority origins (Afghanistan, Cuba, Haiti, Honduras, Venezuela, among others).

**Black Swan Events.** Genuinely novel shocks—sudden asylum-law changes, region-scale disasters, unprecedented policy events—lie beyond the model’s anticipatory reach. The system should complement, not replace, domain expert judgement. Five additional limitations are discussed in Appendix ??.

### 4.3 Future Work

Key extensions include social media integration for finer-grained intent signals, cross-country transfer learning from data-rich to data-sparse origins, real-time streaming deployment, and causal discovery via Granger causality or structural equation models. See Appendix ?? for the full enumeration.

### 4.4 Code and Data Availability

The full codebase, including data-collection scripts, model implementations, and reproduction notebooks, is publicly available at <https://github.com/ibitec7/migration>. Pre-trained model weights, processed datasets, and intermediate artefacts (embeddings, cluster labels, production outputs) are hosted on Hugging Face at <https://huggingface.co/sdsc2005-migration>.

## 5 Contribution Statement

- **Ibrahim:** Multi-modal data ingestion pipeline (async news scraping of 170K+ articles, Trends collection, HF data sync, data processing).
- **Haseeb:** LSTM and Transformer architectures, SurgeJointLoss implementation, model performance visualisations.
- **Asadohoo:** News scraping, URL decoding pipeline, NLP processing (embeddings, clustering, headline scoring).
- **Sameer:** Exploratory data analysis on encounter and visa data (Zipfian analysis, seasonal decomposition), Introduction drafting.
- **Ayaan:** Horizon-Aware Ensemble, per-horizon surge metrics, lead-lag signal analysis across modalities.
- **Pak Hei Ng:** cuML Random Forest baseline, GPU-accelerated clustering, report aggregation, formatting, and references.

## References

- Adrien Barbaresi. Tafilatura: A web scraping library and command-line tool for text discovery and extraction. <https://trafilatura.readthedocs.io>, 2021.
- Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.
- Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 160–172. Springer, 2013.
- Google. Google news RSS feed. <https://news.google.com>, 2025a. Accessed via pygoogle-news RSS API.
- Google. Google trends. <https://trends.google.com>, 2025b. Accessed via pytrends API.
- Michael Günther, Louis Milliken, Jonathan Geuter, Georgios Mastrapas, Bo Wang, and Han Xiao. Jina embeddings: A novel set of high-performance sentence embedding models. *arXiv preprint arXiv:2307.11224*, 2024.
- James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.
- International Monetary Fund. Real effective exchange rate (REER). <https://data.imf.org>, 2025. Accessed via IMF Data portal.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- NVIDIA Corporation. TensorRT: High-performance deep learning inference optimizer and runtime. <https://developer.nvidia.com/tensorrt>, 2024.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. <https://pytorch.org>, 2019.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- RAPIDS Development Team. RAPIDS cuML: GPU-accelerated machine learning. <https://rapids.ai>, 2024.
- Christopher A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980.
- United Nations Department of Economic and Social Affairs. International migration 2020 highlights. Report ST/ESA/SER.A/404, United Nations, 2020.

- U.S. Customs and Border Protection. Southwest land border encounters. <https://www.cbp.gov/newsroom/stats/southwest-land-border-encounters>, 2026. Fiscal Years 2019–2026 (year-to-date).
- U.S. Department of State, Bureau of Consular Affairs. Monthly immigrant visa issuance statistics. <https://travel.state.gov/content/travel/en/legal/visa-law0/visa-statistics/immigrant-visa-statistics/monthly-immigrant-visa-issuances.html>, 2025. Accessed: January 2017 – September 2025.
- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.
- Ritchie Vink. Polars: Blazingly fast DataFrames in Rust and Python. <https://pola.rs>, 2024.

# Appendices

## A Per-Horizon Surge Classification

The summary Table ?? in the main body reports averages across horizons. Tables ??–?? below provide the full break-down for each model on the OOT test set ( $\sim 103$ – $114$  surge instances, threshold  $> 1.5\sigma$ ). The consistent decline in metrics from Lead 1 to Lead 6 reflects the difficulty of long-range forecasting; the ensemble mitigates this decay more effectively than any individual model.

**Table 3:** Random Forest surge classification (OOT).

| Horizon | Prec. | Rec. | F1          |
|---------|-------|------|-------------|
| Lead 1  | 0.96  | 0.97 | <b>0.97</b> |
| Lead 2  | 0.93  | 0.95 | <b>0.94</b> |
| Lead 3  | 0.88  | 0.96 | <b>0.92</b> |
| Lead 4  | 0.85  | 0.94 | <b>0.89</b> |
| Lead 5  | 0.82  | 0.93 | <b>0.88</b> |
| Lead 6  | 0.78  | 0.91 | <b>0.84</b> |

**Table 4:** Transformer Encoder surge classification (OOT).

| Horizon | Prec. | Rec. | F1          |
|---------|-------|------|-------------|
| Lead 1  | 0.95  | 0.93 | <b>0.94</b> |
| Lead 2  | 0.95  | 0.94 | <b>0.94</b> |
| Lead 3  | 0.91  | 0.93 | <b>0.92</b> |
| Lead 4  | 0.86  | 0.90 | <b>0.88</b> |
| Lead 5  | 0.85  | 0.90 | <b>0.88</b> |
| Lead 6  | 0.83  | 0.90 | <b>0.87</b> |

**Table 5:** LSTM (SurgeJointLoss) surge classification (OOT).

| Horizon | Prec. | Rec. | F1          |
|---------|-------|------|-------------|
| Lead 1  | 0.97  | 0.85 | <b>0.91</b> |
| Lead 2  | 0.94  | 0.84 | <b>0.89</b> |
| Lead 3  | 0.91  | 0.83 | <b>0.87</b> |
| Lead 4  | 0.88  | 0.81 | <b>0.84</b> |
| Lead 5  | 0.85  | 0.80 | <b>0.82</b> |
| Lead 6  | 0.82  | 0.79 | <b>0.80</b> |

## B Detailed Event Signal Analysis

Of 392 country–event–lag combinations screened, 58 (14.8%) survived Benjamini–Hochberg correction at  $q \leq 0.10$ . Selected highlights across modalities are described below.

**News event clusters.** Among cluster-level signals, Mexico’s geopolitical-watch cluster showed  $r = -0.436$  at a 1-month lag ( $q = 0.0009$ ), India’s economic-sentiment cluster showed  $r = -0.424$  at 2 months ( $q = 0.0010$ , surge lift 0.932), Honduras’s Evangelical-pastoral cluster showed  $r = +0.327$  at 2 months ( $q = 0.0114$ , lift 1.229), Vietnam’s misinformation cluster

**Table 6:** Horizon-Aware Ensemble surge classification (OOT).

| Horizon | Prec. | Rec. | F1          |
|---------|-------|------|-------------|
| Lead 1  | 0.96  | 0.96 | <b>0.96</b> |
| Lead 2  | 0.93  | 0.96 | <b>0.95</b> |
| Lead 3  | 0.92  | 0.94 | <b>0.93</b> |
| Lead 4  | 0.88  | 0.94 | <b>0.91</b> |
| Lead 5  | 0.83  | 0.94 | <b>0.88</b> |
| Lead 6  | 0.80  | 0.92 | <b>0.86</b> |

showed  $r = -0.403$  at 5 months ( $q = 0.0017$ , lift 0.816), and El Salvador’s conflict cluster showed  $r = -0.361$  at 5 months ( $q = 0.0055$ , lift 0.842). Eleven of the 15 countries exhibited at least one significant event signal.

**Headline sentiment scores.** Aggregate headline-level scores revealed sharper relationships still. The Dominican Republic produced the single strongest lead in the study:  $r = 0.617$  at a 3-month lag ( $q = 3.3 \times 10^{-9}$ ). Cuba’s “Trump Policy” cluster showed  $r = -0.595$  at lag 0 ( $q = 3.4 \times 10^{-9}$ ), capturing the immediate visa impact of U.S. policy shifts.

**Exchange rate and Trends signals.** Among the 6 countries with REER data, the Dominican Republic again dominated ( $r = 0.498$ , lag 2,  $q = 2.6 \times 10^{-6}$ ), followed by Mexico ( $r = 0.289$ , lag 6,  $q = 0.018$ ). Google Trends was significant for 66 of 120 rows (55.0%); notable examples include China’s `us_visa` keyword ( $r = 0.406$ ,  $q = 1.1 \times 10^{-4}$ ) and Afghanistan’s `us_jobs` ( $r = -0.484$ ,  $q = 4.4 \times 10^{-6}$ ).

## C Extended Limitations

The following supplement the four limitations discussed in the Conclusion (§4.2).

**Google Trends as a Weak Standalone Predictor.** Despite significant keyword-level correlations, the trends-enhanced VAR model underperformed the autoregressive baseline (−5.97% RMSE on average), indicating that search-intent signals require non-linear architectures to be leveraged effectively.

**News Coverage Bias.** The 61.1% pipeline success rate means roughly 39% of articles were lost to extraction failures, paywalls, or anti-scraping measures. Coverage is skewed toward English-language media, potentially underrepresenting local-language reporting.

**Feature Interpretability and Selective Collection.** Random Forest feature importance scores are currently unused for guiding data collection. Systematic importance analysis could concentrate future efforts on high-signal inputs, reducing operational cost.

**Model Staleness and Distributional Drift.** The model is anchored to the 2017–2022 training window. Concept drift—particularly shifts in the Zipfian distribution of visa issuances as geopolitical pressures evolve—could push the system out-of-distribution without periodic retraining.

**Survivorship Bias in Encounter Data.** Border encounters record only *detected* crossings; migrants who evade apprehension are absent. The model therefore partially learns enforcement detection rates rather than true migration pressure.

## D Ensemble Weight Schedule

Table ?? shows the horizon-specific weighting scheme.

**Table 7:** Horizon-Aware Ensemble weights (sum to 1.0 per row).

| Lead | RF  | Trans. | LSTM | Rationale                          |
|------|-----|--------|------|------------------------------------|
| 1 mo | 60% | 10%    | 30%  | RF & LSTM dominate short-term      |
| 2 mo | 50% | 20%    | 30%  | RF still dominant                  |
| 3 mo | 30% | 40%    | 30%  | Balanced transition                |
| 4 mo | 20% | 60%    | 20%  | Transformer increasingly critical  |
| 5 mo | 10% | 80%    | 10%  | Transformer dominates long-horizon |
| 6 mo | 5%  | 85%    | 10%  | Transformer captures slow trends   |

## E Future Work

1. **Social media integration:** Signals from X (Twitter), Facebook, and WhatsApp could provide finer-grained, less media-filtered insight into migration intent.
2. **Cross-country transfer learning:** Training on data-rich countries (Mexico, India) and transferring representations to data-sparse origins (Haiti, Afghanistan).
3. **Real-time streaming:** Adapting the batch-monthly pipeline for continuous ingestion with rolling daily or weekly forecasts.
4. **Richer macro indicators:** Incorporating inflation, unemployment, remittance flows, and commodity prices to fill exchange-rate gaps.
5. **Causal discovery:** Applying Granger causality, structural equation models, or natural-experiment designs to move beyond predictive correlation.

## F Data Pipeline Engineering Details

The news collection pipeline was engineered for throughput and fault tolerance:

- Articles were collected via the Google News RSS API (?), with encoded URLs resolved through a batched decoding endpoint (20 articles per batch), reducing API calls by ~95%.
- Full article text was fetched asynchronously with semaphore-controlled concurrency (max 16 connections), exponential backoff with jitter, and a 30-second watchdog timeout per request.
- HTML-to-text extraction was performed via Trafilatura (?), parallelized across CPU cores using a `ProcessPoolExecutor` (4 workers).
- Intermediate results were checkpointed every 25 articles via atomic writes (write to temp file, then rename), ensuring crash-resilient resumption.

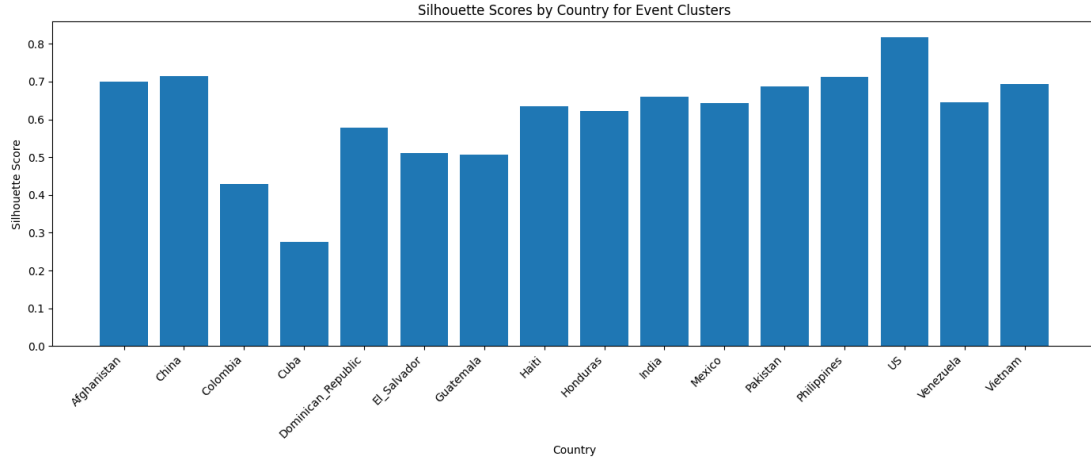
Exploratory and model-performance plots (visa time series, encounter trends, cluster UMAP visualisations, F1 trajectory charts, and lead-lag heatmaps) are stored in the `data/plots/` directory of the project repository, organised into sub-folders: `eda_plots/`, `cluster_plots/`, `model_performance/`, and `trends_vs_migration/`.

## G Cluster Quality Analysis

Figures ?? and ?? summarise the quality and thematic granularity of the HDBSCAN event clusters produced for each of the 15 origin countries.

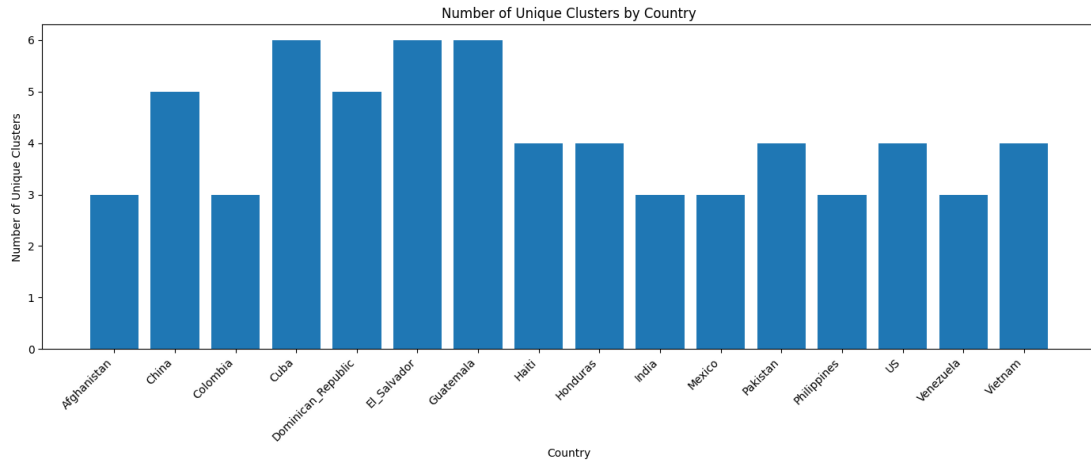
**Silhouette Scores.** A silhouette score above 0.5 indicates well-separated, internally coherent clusters. As Figure ?? shows, the majority of countries achieve scores above this threshold; the United States (US-centric context articles), China, and the Philippines reach 0.70–0.82, confirming that Jina v5 embeddings capture thematically distinct news groupings. Cuba and Colombia

are the notable exceptions (scores 0.28–0.43), reflecting noisier or more topically overlapping news coverage for those origins.



**Figure 3:** Silhouette scores by country for HDBSCAN event clusters. Values above 0.5 indicate well-separated, internally coherent clusters. Most countries exceed this threshold, validating the Jina v5 embedding space as a meaningful basis for thematic grouping.

**Number of Unique Clusters.** Figure ?? shows that Cuba, El Salvador, and Guatemala produced the largest number of distinct event themes (6 clusters each), reflecting richer and more varied news landscapes for these countries in the context of U.S. migration. Countries with fewer clusters (3) such as Afghanistan, Colombia, and India tended to have more concentrated coverage around a single dominant narrative (e.g., Taliban governance or H-1B visa policy), which can make cluster labels simpler but potentially reduces signal diversity.

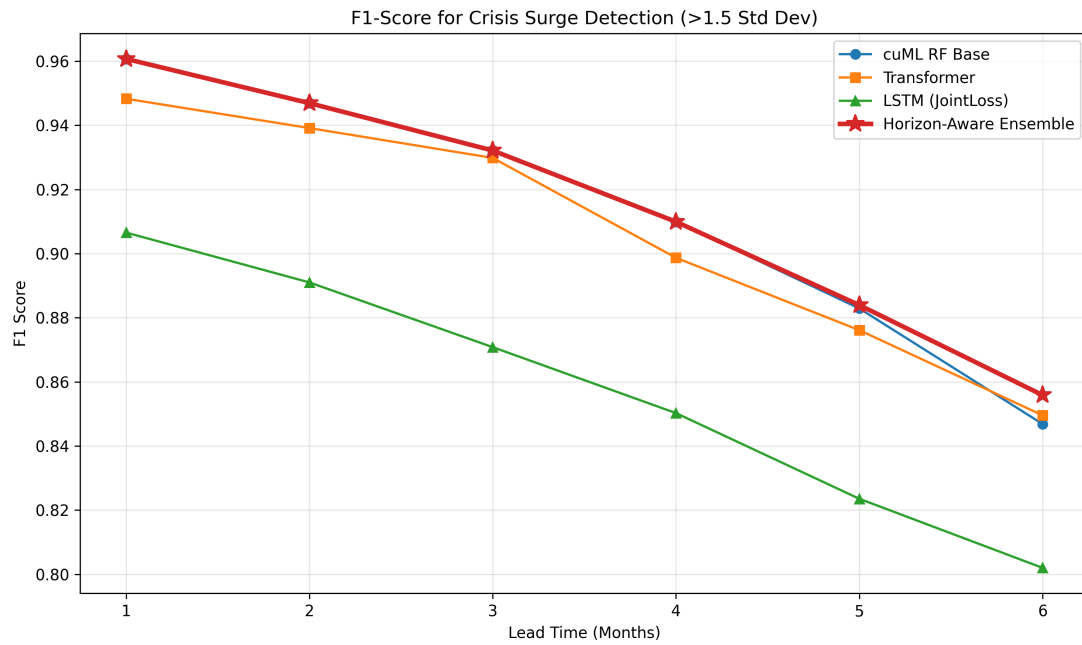


**Figure 4:** Number of unique HDBSCAN event clusters per country. Greater cluster counts indicate richer thematic diversity in the collected news corpus, providing more candidate leading indicators per origin.

## H F1-Score Trajectory Across Forecast Horizons

Figure ?? plots per-lead F1 scores for all four models on the OOT test set, complementing the tabular summaries in Tables ??–?? and giving an at-a-glance view of how each architecture’s surge-detection ability degrades with forecast distance.

The ensemble curve is notably flatter than that of any individual model, confirming that the horizon-specific weighting schedule (Appendix ??) successfully compensates for the comple-



**Figure 5:** F1 scores for surge detection across forecast lead times (Leads 1–6). All models degrade with increasing horizon; the Horizon-Aware Ensemble (red star) consistently outperforms individual architectures at every lead. The LSTM (JointLoss, green triangle) trades lower F1 for higher precision, visible as the most conservative trajectory.

mentary decay profiles of Random Forest and Transformer. The Horizon-Aware Ensemble maintains  $F1 > 0.86$  even at Lead 6—a 2-percentage-point improvement over the best single model at that horizon.