# Forecasting and Nowcasting with Text as Data

## Module 2 - Session 2: From signals to decisions

Renato Vassallo

May, 2026

Barcelona School of Economics

*All views expressed here and any remaining errors are my own.*

# Introduction

# From zero-shot to few-shot

- Zero-shot works well when labels are well described in natural language.
- In many applications, tasks are:
  - domain-specific,
  - ambiguous,
  - or poorly captured by generic labels.
- We often have **a small number of labeled examples**.

*How can we learn effectively from very limited supervision?*

# Medicine

- Detect rare conditions with very limited labeled data
- Only a few annotated scans are available
- The model learns patterns and generalizes to new patients

*Same idea: learn from similarity rather than large datasets*

# Medicine

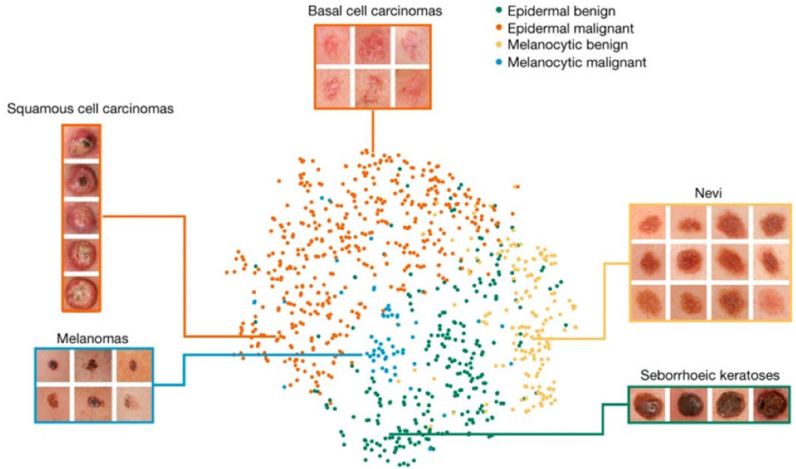



**Figure 1:** CNN internal representations of four key disease classes using t-SNE on the last hidden layer (932 biopsy-proven images). Source: Esteva et al. (2017).

# Face recognition

- Real-world setting: identify individuals with very few images
- Example: only 2–3 photos per person (e.g., security, surveillance)
- Model compares new images to known identities in an embedding space

Recent work: Open-set face recognition (2023)

# Face recognition


**Figure 2:** Attention visualization of different expressions. In each set of examples, the first row represents the original image and the second row represents the generated class activation map. Source: Chen et al. (2023).

# Few-shot learning

# Core idea

- Learn a **representation space** where similar inputs are close.
- Use a small labeled set to define the task.

$$\mathcal{S} = \{(d_i, y_i)\}_{i=1}^{m}, \quad m \ll N$$

- Generic formulation:

$$\hat{y}_i = g_{\phi(\mathcal{S})}(f_\theta(d_i))$$

- The model adapts using $\mathcal{S}$, not large datasets.

# Why few-shot in social science?

- Many tasks have:
  - limited labeled data
  - high annotation costs
  - evolving definitions
- Examples:
  - conflict events
  - economic sentiment
  - institutional disruption ▸ Predicting Autocratization

*Few-shot learning allows task-specific adaptation at low cost*

# SetFit: efficient few-shot learning

Developed by Tunstall et al. (2022), SetFit is a practical two-step few-shot approach

▸ SetFit Architecture

**Step 1:** fine-tune the embedding space using contrastive learning

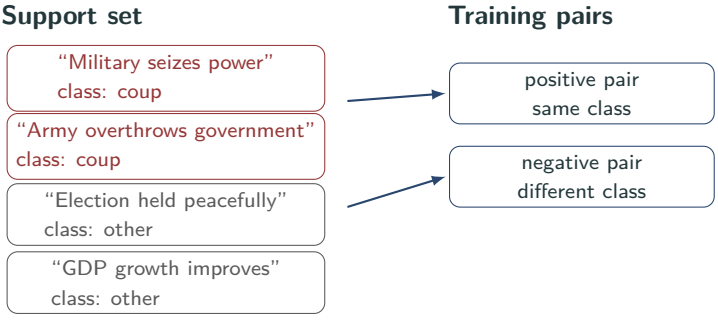
$$f_{\theta_0} \rightarrow f_{\hat{\theta}}$$

**Step 2:** train a lightweight classifier on top

$$\hat{y}_i = g_{\hat{\phi}}(f_{\hat{\theta}}(d_i))$$

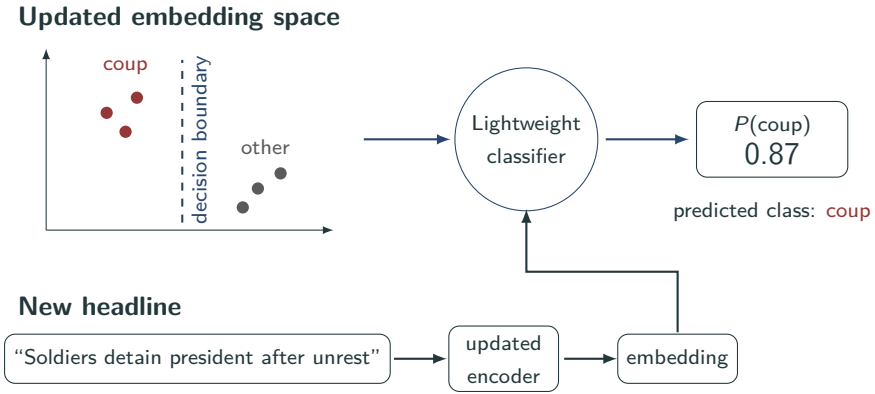
- Works well with very small labeled datasets
- Fast and computationally efficient
- Often stronger than naive fine-tuning in low-data settings

# Fine-tuning the embedding space




SetFit first learns a task-specific embedding space by pulling same-class pairs together and pushing different-class pairs apart.

# Classifier on updated embeddings




After adapting the embedding space, SetFit trains a lightweight classifier that maps updated embeddings into class probabilities.

# From theory to practice

- We now apply few-shot learning to:
  - institutional disruption
  - de-democratization events
- Using SetFit and a small labeled dataset
- Before moving on, let's quickly recall evaluation metrics ▸ Precision–Recall

An empirical illustration is provided in `session2/01_signals_to_decisions.ipynb`

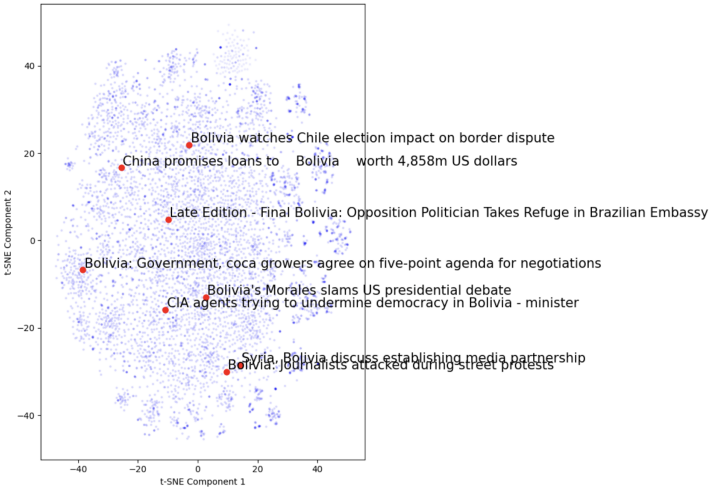
# Institutional disruption

# Predicting autocratization: main take-aways

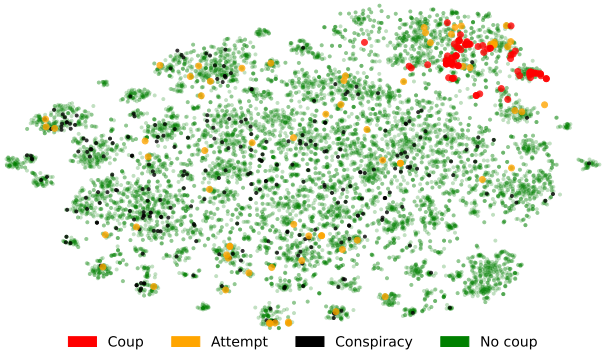
- Mayoral et al. (2026): nowcasting and forecasting de-democratization events using text.
- Goal: real-time monitoring of institutional instability.
- Very difficult task (needle in the haystack)
- Monthly updates for risk are possible
  - It works for coups! (Cline included several of our nowcasts).
  - Good results for term limits
  - But: human coding will remain important
- Some gains from adding more events
- Points towards a measure of institutional *fragility*

# Embeddings and few-shot learning

- We reduce 6 million newspaper headlines into embeddings



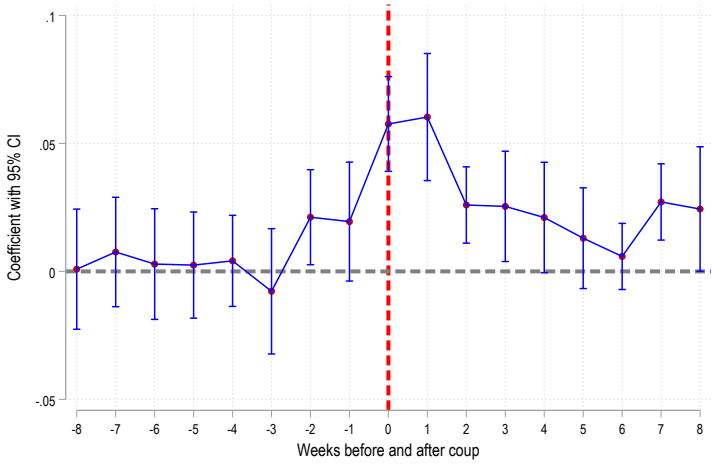
# Newspaper headlines of DRC



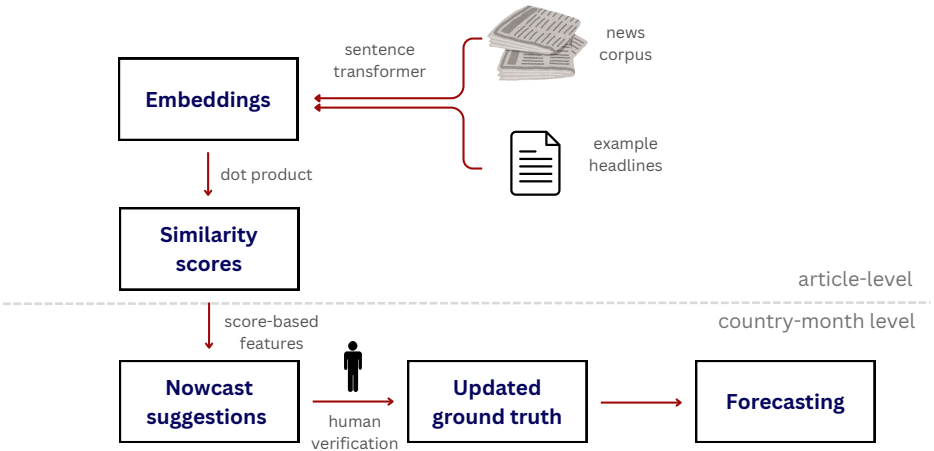

- Red dots are around when coups happen
- Embeddings of headlines during coups are clustered

# Mean few-shot similarity score

Few-shot similarity scores shoots up after coup (and actually already before).

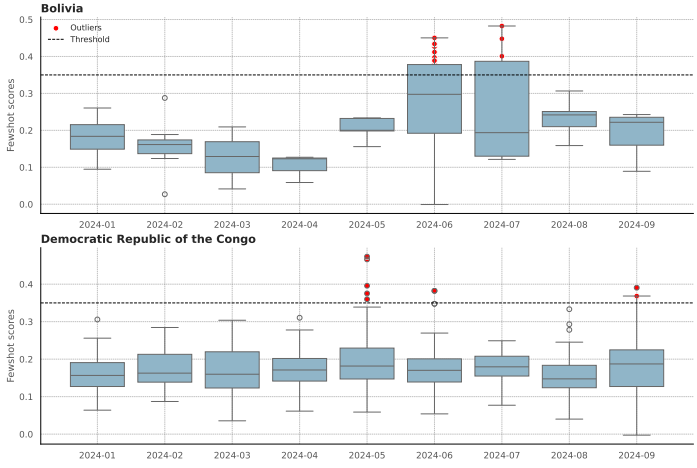


# System architecture



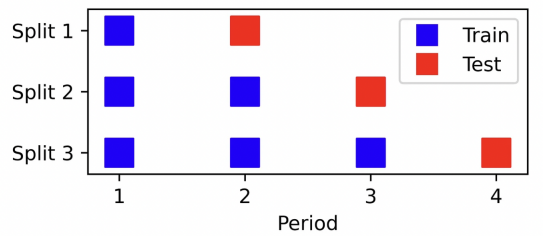
# Nowcast for target updates

Case studies of distributions of similarity scores around events



# Rolling forecast

- Use data until today to predict tomorrow: pseudo out-of-sample evaluation.
- Model: tree-based (Random Forest and XGBoost).
- Feature space:
  1. Historical data
  2. Text features (LDA)
  3. Political indices (VDEM)
  4. Socioeconomic indicators (WB)



# Combining Event Types: Same Train and Test

| Train data | Test data | ROC-AUC | | Avg Precision |
|---|---|---|---|---|
| | | Overall | Hard | |
| Coup | Coup | 0.72 | 0.77 | 0.10 |
| Coup + TLE | Coup + TLE | 0.77 | 0.78 | 0.14 |
| Coup + TLE + JW | Coup + TLE + JW | 0.76 | 0.72 | 0.29 |

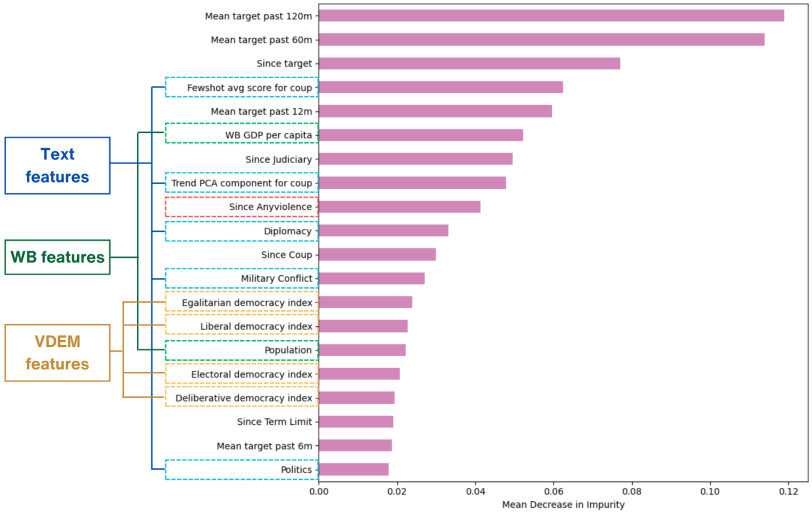
**Note:** Performance metrics for institutional disruption forecast, 12 months ahead. Rolling forecast strategy from Jan 2010 until Mar 2025 using Random Forest Classifier. **Coup**: Coup d'État events, **TLE**: term-limit evasion events, and **JW**: judiciary weakened events. Each row aggregates events to the previous.

# Combining Event Types: Expanded Training Set

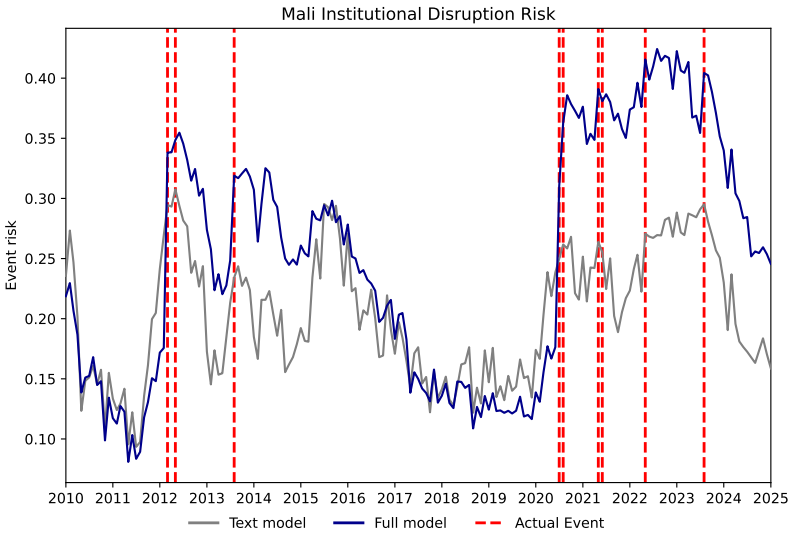
| Train data | Test data (with NaNs) | ROC-AUC | | Avg Precision |
|---|---|---|---|---|
| | | Overall | Hard | |
| Coup | Coup | 0.76 | 0.76 | 0.11 |
| Coup + TLE + JW | Coup | 0.77 | 0.76 | 0.11 |
| TLE | TLE | 0.78 | 0.76 | 0.07 |
| TLE + Coup | TLE | 0.80 | 0.77 | 0.08 |
| JW | JW | 0.78 | 0.68 | 0.22 |
| JW + TLE | JW | 0.79 | 0.71 | 0.23 |

**Note:** Performance metrics for institutional disruption forecast, 12 months ahead. Rolling forecast strategy from Jan 2010 to Mar 2025 using Random Forest Classifier. **Coup**: Coup d'État events, **TLE**: term-limit evasion events, and **JW**: judiciary weakened events.

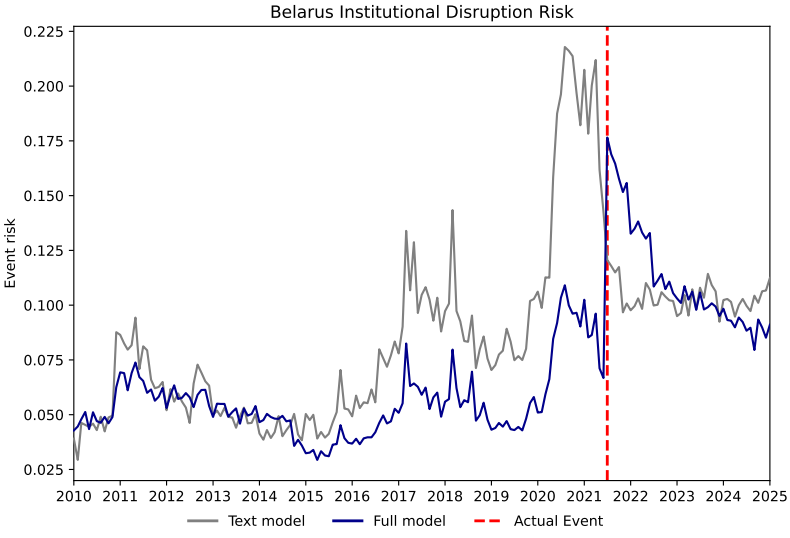
# Feature importance for full RF model



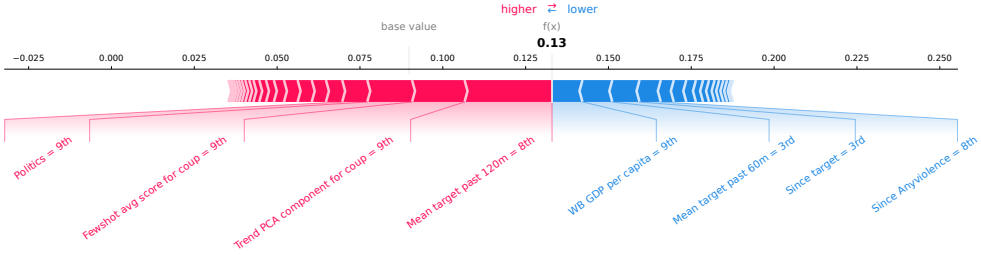
# Forecasting institutional risk: Mali



# Forecasting institutional risk: Belarus



# Shapley values



# Local projections to validate event data

Local projections diff-in-diff specification (Jordà (2005) and Dube et al. (2023)).

$$
\begin{aligned}
y_{c,t+h} - y_{c,t-1} = \quad & \beta_h^{LP-DiD} \Delta T_{ct} && \} \text{ treatment indicator} \\
& + \sum_{j=1}^{p} \gamma_j^h y_{c,t-j} && \} \text{ outcome lags} \\
& + \delta_t^h && \} \text{ time effects} \\
& + e_{ct}^h && \text{For } h = 0, \ldots, H.
\end{aligned}
$$

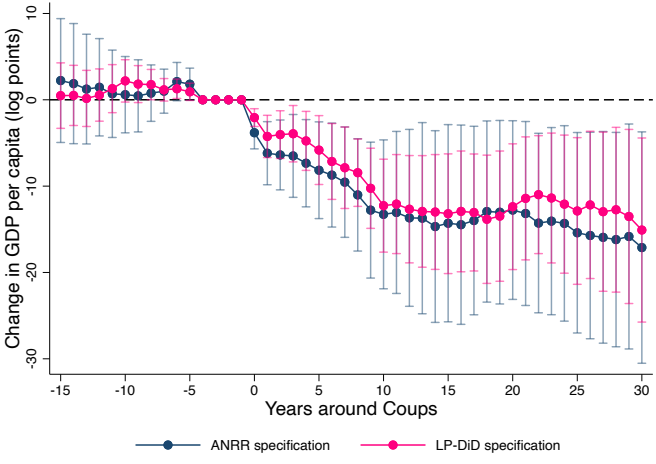
Restricting the estimation sample to:

$$
\begin{cases}
\text{events} & T_{ct} = 1;\ T_{c,t-1} = 0 \\
\text{outcome lags} & \text{We set } p = 4 \text{ as in Acemoglu et al. (2019).}
\end{cases}
$$

# Impact study

back

Treatment effects of Coups d'Etat on the log of GDP per capita



# Crime Index

# Measuring the perception of crime from global news

- Ongoing work with IMF.
- Goal: transform global news into high-frequency measures of **crime perception**.
- $\sim$20 million articles (Jan 2006 – Apr 2025)
- 16 countries in Latin America and the Caribbean
- Global and local news outlets
- Spanish ($\sim$60%) and English ($\sim$40%)

# Pipeline

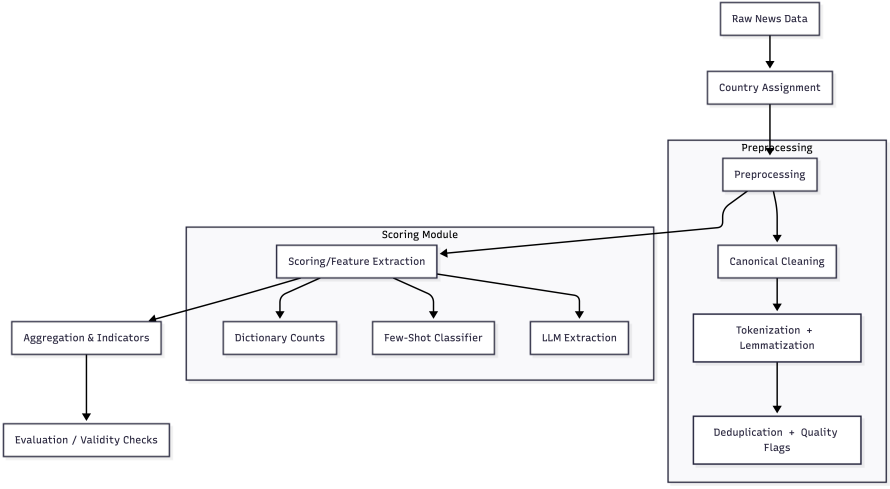
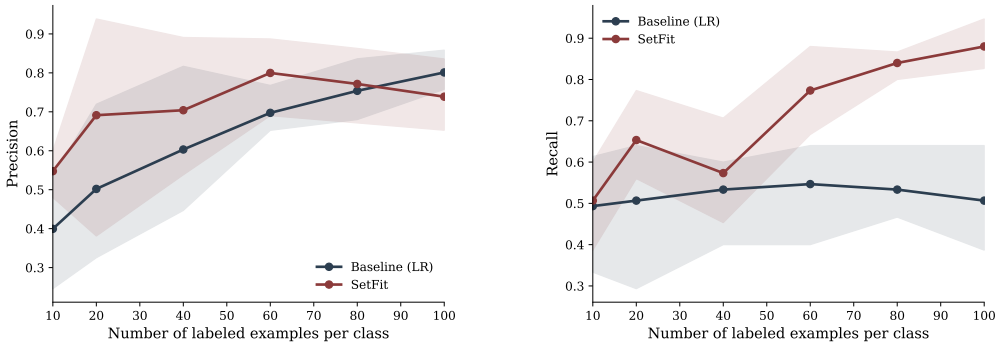



**Figure 3:** News-based crime indicator construction

# Model-based approach

- Dictionary methods can be noisy: limited context and fixed vocabularies.
- We use a model-based classifier with SetFit.
- **Ground-truth dataset**: 1,000 labeled articles (200 crime, 800 non-crime), including *hard negatives* to reduce false positives.
- **Few-shot evaluation**: repeatedly sample training subsets to assess performance and sensitivity to the number of labels.

# Model-based approach




**(a)** Precision **(b)** Recall

**Figure 4:** Learning curves under precision–recall constrained threshold

*Notes:* The figure shows out-of-sample precision and recall for a baseline LR classifier and a SetFit model as the number of labeled examples per class increases. Thresholds are chosen on a calibration set by maximizing precision subject to recall $\geq 0.60$. Points denote mean performance across random splits; shaded areas indicate 10th–90th percentile ranges.

# Country case: El Salvador

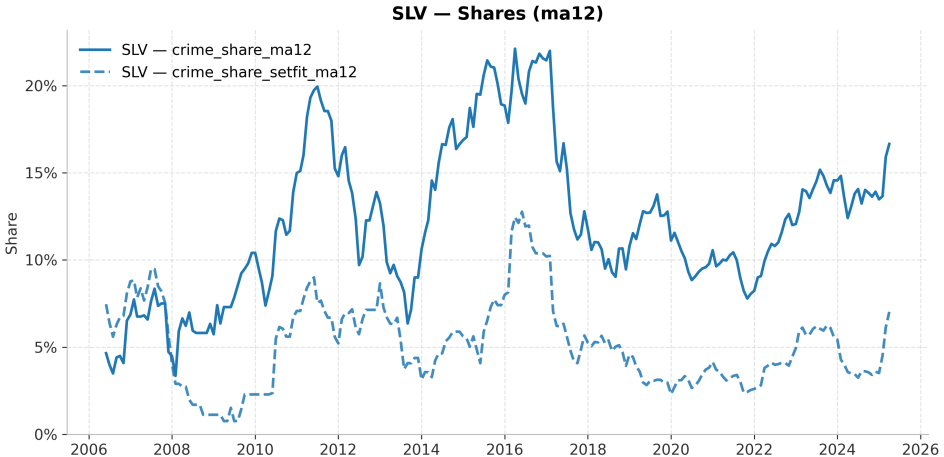
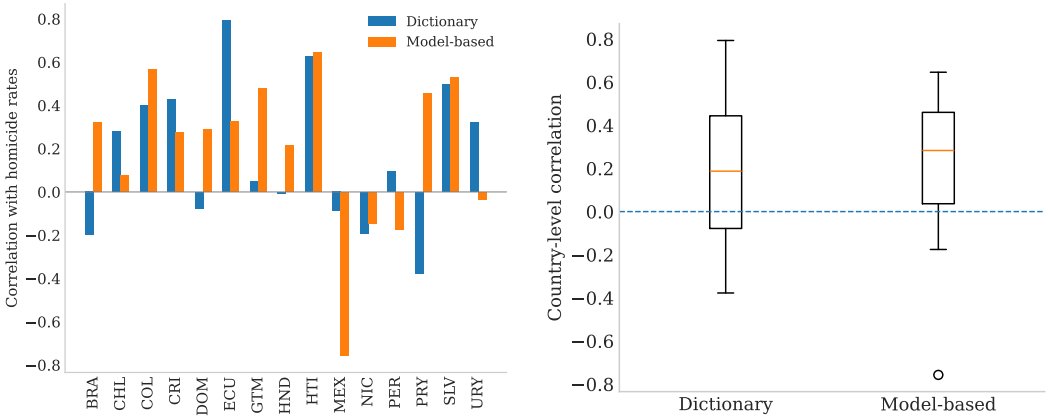



**Figure 5:** Dictionary and model-based crime index for El Salvador (12-months moving average)

# Example: Dictionary vs Model-Based Classification

| title | crime_dict | setfit_prob |
|---|---|---|
| Suspect in 4 slayings accused of selling victims' jewelry | 1 | 0.99678 |
| Gang leader captured who in 2008 was acquitted for murder | 1 | 0.98449 |
| El Salvador launches new arrest order against former president Funes | 1 | 0.443295 |
| Against corruption, strong institutions | 1 | 0.307239 |
| Salvadoreños buscan un nuevo presidente que atienda inseguridad y economía | 1 | 0.0525 |
| Central Americans yearning for U.S. turn to smugglers amid Trump asylum crackdown | 1 | 0.03117 |
| Illegal Migrant Crossings Surge in Remote New Mexico Desert | 0 | 0.99549 |
| Wall-weary U.S. Republicans pivot toward immigrant deportations | 0 | 0.99578 |
| Un motociclista fallecido a diario en percances viales | 0 | 0.99605 |

# Validation: correlation with actual homicides




**(a)** Correlations by country

**(b)** Distribution of country correlations by method

**Figure 6:** Correlation of news-based crime indicators with homicide rates. Source: UNODC.

# Crime perception and industrial production

**Question**

How do changes in crime perception affect economic activity?

- Build a panel of LAC countries with:
  - Industrial production.
  - Key macro variables: prices, interest rates, exchange rates.
  - News-based crime indicator.
  - US industrial production (common exogenous factor).
- Estimate a Bayesian panel VAR to trace impulse responses:

$$y_{i,t} = A_1 y_{i,t-1} + \cdots + A_{12} y_{i,t-12} + B x_{i,t} + u_{i,t}, \qquad u_{i,t} \sim \mathcal{N}(0, \Sigma)$$

- Recursive Identification: IP $\rightarrow$ CPI $\rightarrow$ Policy Rate $\rightarrow$ Exchange Rate $\rightarrow$ Crime. Crime last $\rightarrow$ innovations influence the economy only with a delay.

# Impact study

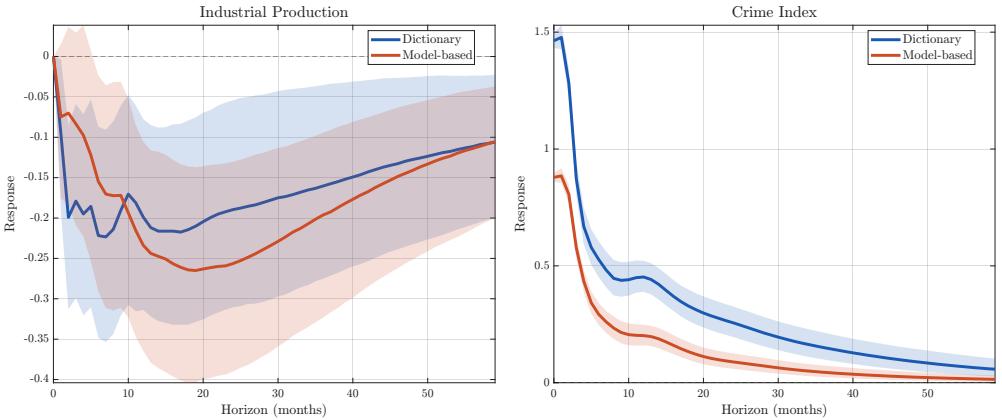



**Figure 7:** Impulse responses to a 1 s.d. shock in crime indices. The model is a Bayesian Panel VAR with a Minnesota-type prior, including key macroeconomic variables as controls; crime indices are ordered last under recursive identification.

# LLM approach

- Starting point: dictionary-based indicator flags potentially relevant articles.
- For each flagged article, we query an LLM[1] to:
  - Classify the type (violent vs white-collar vs unclear vs not-crime).
  - Additional features (tense, country match, etc).
- Outcome:

| isocode | title | category | tense | country_match |
|---|---|---|---|---|
| PER | Mexico unions ready for stoppages; conflict worsens | violent | past | FALSE |
| PER | Peru's new leader must attack waste, corruption | white_collar | unclear | TRUE |
| PER | Ataques priman en debate candidatos presidencia Perú | unclear | unclear | TRUE |

[1]We employ Llama 3.1 (8B parameters), consistent with our computational resources and with the implementation strategy in Braun and Oswald (2025). ▸ Benchmarking LLMs

# LLM refinement: white-collar crime

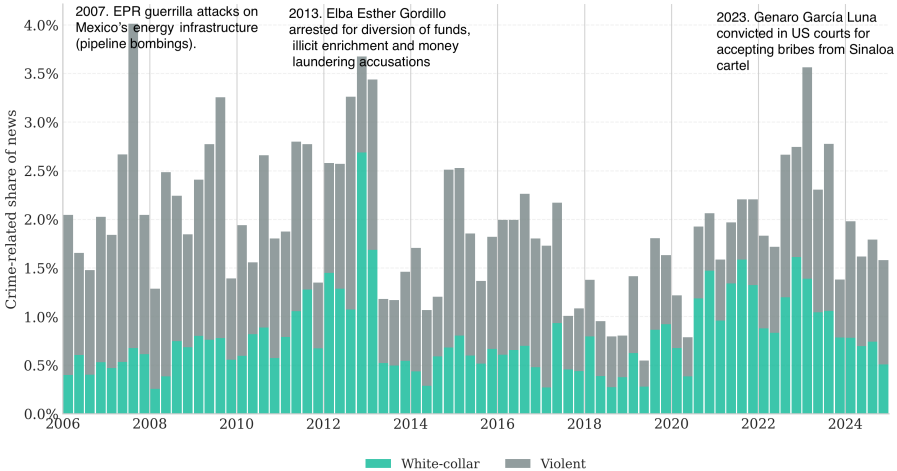



**Figure 8:** Crime index for Mexico, disaggregated into violent and white-collar components

# Forecasting white-collar scandals

**Question**

Can the white-collar index anticipate large scandals?

- Setup:
  - Focus on Mexico and Peru.
  - Define "big scandals" as months where the white-collar index exceeds 2 standard deviations above its mean.
  - Use lagged values of the white-collar index (and other controls) to forecast these events.
- Rolling expanding-window forecast strategy.

# Forecasting white-collar scandals

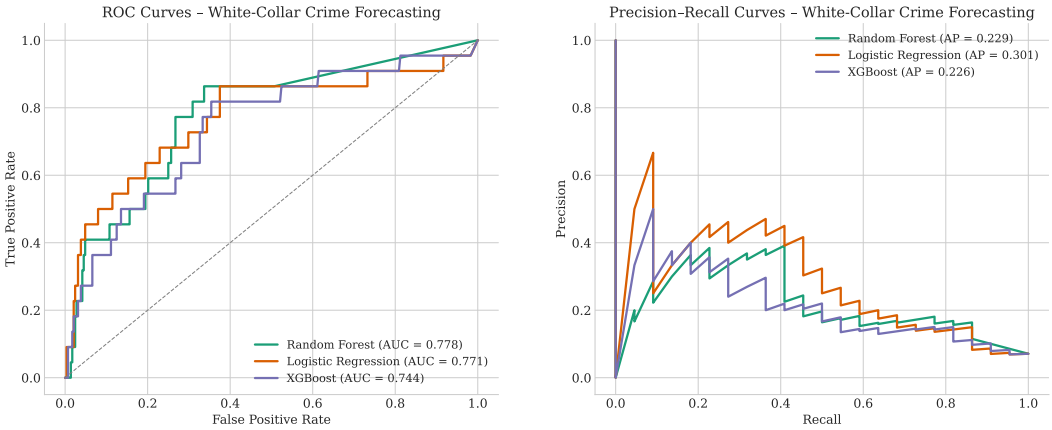



**Figure 9:** Model performance for predicting white-collar crime events.

# Appendix

# Precision vs. recall: coup detector

back

$$\hat{y}_i = \mathbb{1}(\hat{p}_i \geq 0.5)$$

| # | True | Prob. | Pred. |
|---|---|---|---|
| 1 | C | 0.91 | |
| 2 | C | 0.82 | |
| 3 | C | 0.76 | |
| 4 | C | 0.64 | |
| 5 | C | 0.32 | |
| 6 | N | 0.21 | |
| 7 | N | 0.62 | |
| 8 | N | 0.53 | |
| 9 | N | 0.51 | |
| 10 | N | 0.07 | |

# Precision vs. recall: coup detector

back

$$\hat{y}_i = \mathbb{1}(\hat{p}_i \geq 0.5)$$

| # | True | Prob. | Pred. |
|---|---|---|---|
| 1 | C | 0.91 | C |
| 2 | C | 0.82 | C |
| 3 | C | 0.76 | C |
| 4 | C | 0.64 | C |
| 5 | C | 0.32 | N |
| 6 | N | 0.21 | N |
| 7 | N | 0.62 | C |
| 8 | N | 0.53 | C |
| 9 | N | 0.51 | C |
| 10 | N | 0.07 | N |

# Precision vs. recall: coup detector

back

$$\hat{y}_i = \mathbb{1}(\hat{p}_i \geq 0.5)$$

| # | True | Prob. | Pred. |
|---|---|---|---|
| 1 | C | 0.91 | C |
| 2 | C | 0.82 | C |
| 3 | C | 0.76 | C |
| 4 | C | 0.64 | C |
| 5 | C | 0.32 | N |
| 6 | N | 0.21 | N |
| 7 | N | 0.62 | C |
| 8 | N | 0.53 | C |
| 9 | N | 0.51 | C |
| 10 | N | 0.07 | N |

Then:

- **Precision:** Precision $= \frac{4}{7} \approx 0.57$
  *Many false alarms*

# Precision vs. recall: coup detector

back

$$\hat{y}_i = \mathbb{1}(\hat{p}_i \geq 0.5)$$

| # | True | Prob. | Pred. |
|---|---|---|---|
| 1 | C | 0.91 | C |
| 2 | C | 0.82 | C |
| 3 | C | 0.76 | C |
| 4 | C | 0.64 | C |
| 5 | C | 0.32 | N |
| 6 | N | 0.21 | N |
| 7 | N | 0.62 | C |
| 8 | N | 0.53 | C |
| 9 | N | 0.51 | C |
| 10 | N | 0.07 | N |

Then:

- **Precision:** Precision $= \frac{4}{7} \approx 0.57$
  *Many false alarms*
- **Recall:** Precision $= \frac{4}{5} = 0.80$
  *Most coups detected*

# Precision vs. recall: coup detector

back

$$\hat{y}_i = \mathbb{1}(\hat{p}_i \geq 0.5)$$

| # | True | Prob. | Pred. |
|---|---|---|---|
| 1 | C | 0.91 | C |
| 2 | C | 0.82 | C |
| 3 | C | 0.76 | C |
| 4 | C | 0.64 | C |
| 5 | C | 0.32 | N |
| 6 | N | 0.21 | N |
| 7 | N | 0.62 | C |
| 8 | N | 0.53 | C |
| 9 | N | 0.51 | C |
| 10 | N | 0.07 | N |

Then:

- **Precision:** Precision $= \frac{4}{7} \approx 0.57$
  *Many false alarms*
- **Recall:** Precision $= \frac{4}{5} = 0.80$
  *Most coups detected*

Policy trade-off

- If missing coups is very costly $\Rightarrow$ prioritize **recall**
- If false alarms are costly $\Rightarrow$ prioritize **precision**

# SetFit architecture

back

1. Convert each article into embeddings using a pretrained language model.
2. Fine-tune these embeddings with a small labeled dataset using **contrastive learning**.
3. Train a simple classifier on these updated embeddings.

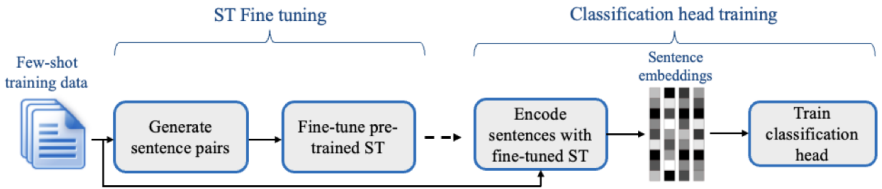



**Figure 10:** SetFit's fine-tuning and training block diagram (Tunstall et al., 2022)

# Braun and Oswald (2025) - Benchmarking LLM Performance

back

- Benchmark 14 LLMs on extracting structured information from conflict event notes; *GPT-4o* and *Llama-3.3-70B* achieve high accuracy at relatively low computational cost.

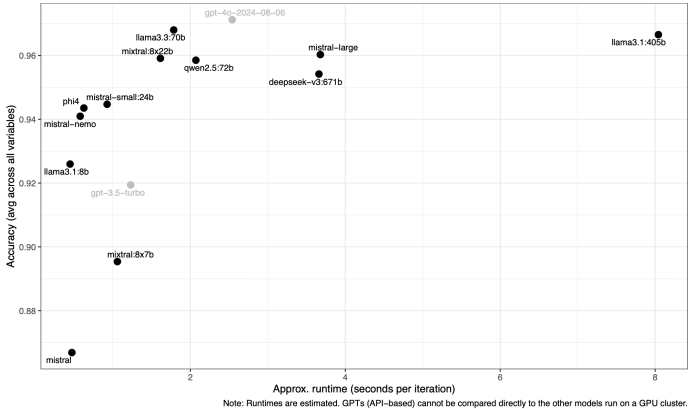



**Figure 11:** Trade-Off between Performance and Computational Costs. Source: Braun and Oswald (2025).

# References

Braun, L. and Oswald, C. (2025). Automated information extraction from text variables in event datasets with large language models. OSF Preprint.

Chen, X., Zheng, X., Sun, K., Liu, W., and Zhang, Y. (2023). Self-supervised vision transformer-based few-shot learning for facial expression recognition. *Information Sciences*, 640:119091.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

Mayoral, L., Mueller, H., Rauh, C., Phillip, M., and Vassallo, R. (2026). Semantic similarity measures in newspaper text for detecting and predicting disruptive institutional events. *BSE Working Paper 1555*. Barcelona School of Economics.

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., and Pereg, O. (2022). Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.