

Early-Warning NLP for Fraud Detection in Consumer Complaints

Milestone 3 — Analysis &
Results

Hasnaa Elidrissi



Business Problem



Fraud patterns evolve faster than playbooks



Analysts spend hours routing incoming complaints manually



Goal: build an early-warning pipeline from opt-in CFPB complaint narratives to
High-precision triage cues, and
Topi-shift alerts signaling new fraud trends

Data & Privacy Controls

Source:

CFPB Consumer
Complaint Database
(Jan 2024 – Sep 2025).

Records:

61,694 unique
complaints; 99 %
include narrative text.

Pre-processing:

deduplication +
HMAC-SHA256
pseudonymization of
emails, phones, digits.

Storage:

Parquet files for
reproducibility.

Weak Label Bootstrapping



Weak supervision via fraud-related keywords.

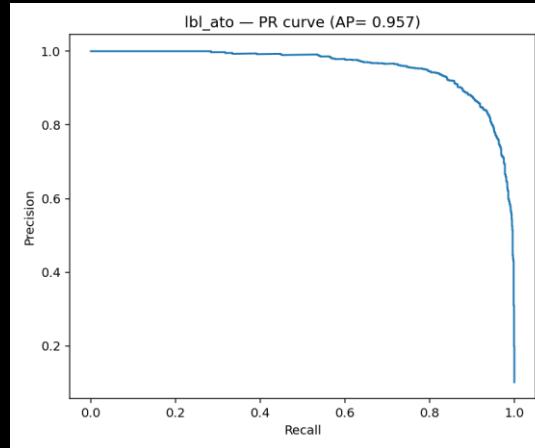
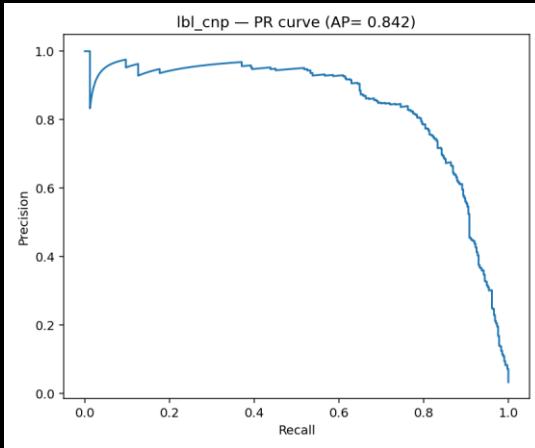
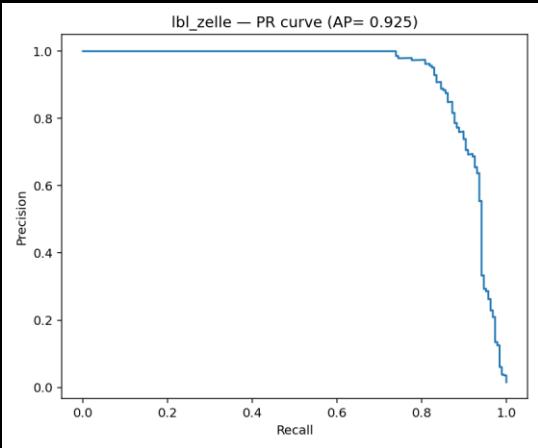
Four labels: ATO, CNP, Zelle/P2P, Phishing.

Guard rule excludes credit-report only cases.

Label Prevalence and Coverage

Label	Prevalence (%)	Positives	Avg Precision	Coverage (%)
ATO	10.2	6 274	0.97	13.4
Zelle/P2P	1.5	942	0.93	1.9
CNP	3.3	2 063	0.72	2.8

Model Performance (PR Curves)



ATO

AP = 0.96, P = 0.75 R = 0.98

Zelle

AP = 0.93, P = 0.75 R = 0.91

CNP

AP = 0.84, P = 0.75 R = 0.62

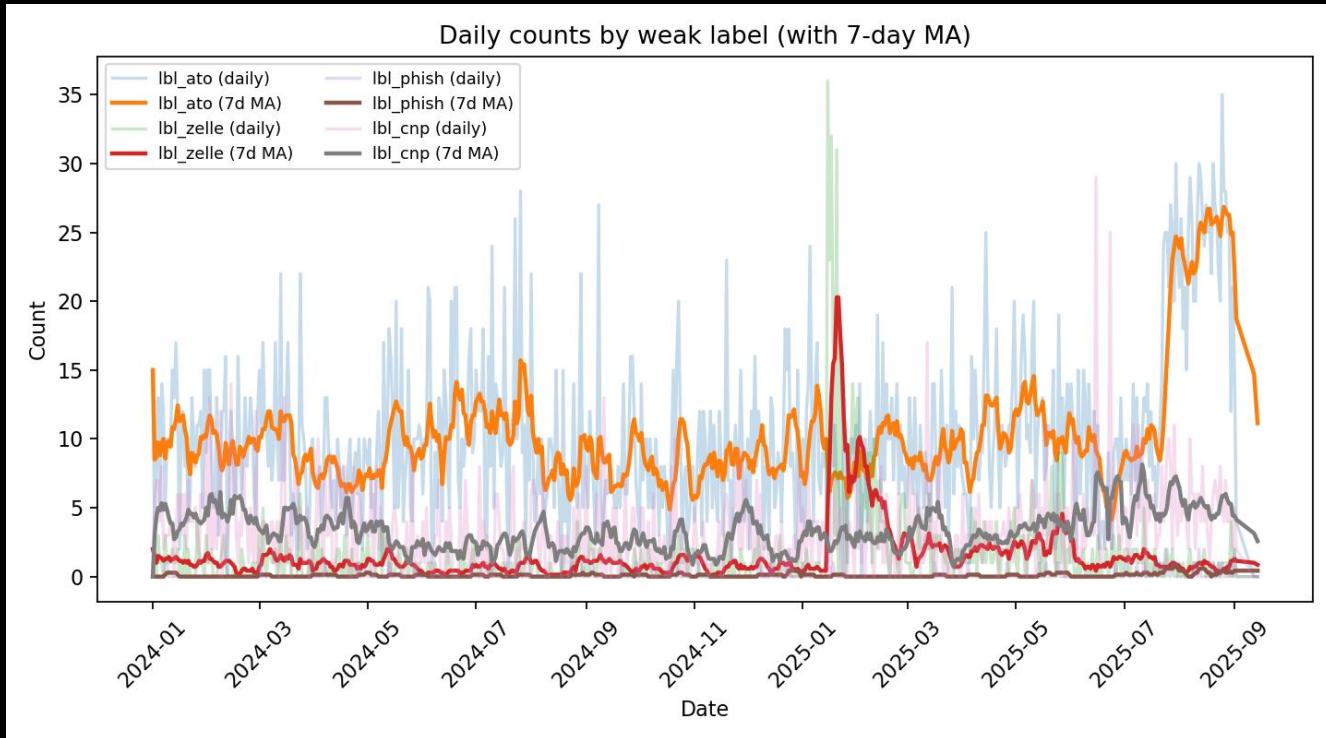
Precision vs Recall Trade-offs

Precision-first thresholds minimize false positives.

Relaxing thresholds increases recall: broader coverage.

Operational sweet spot: $P \geq 0.75$.

Topic Trends & Early Warnings



NMF TOPICS REVEAL SLOWLY EVOLVING THEMES.

SPIKES IN “VERIFICATION CODE,” “ZELLE SENT,” COINCIDE WITH ATO & P2P EVENTS.

DEMONSTRATES FEASIBILITY OF NARRATIVE-BASED EARLY WARNING.

Text-Only Prediction of Monetary Relief

Metric	Class 0 (No Relief)	Class 1 (Relief)	Weighted / Overall
Precision	0.593	0.701	—
Recall	0.958	0.130	—
F1-score	0.733	0.219	—
Accuracy			0.602

- Logistic regression on narrative text predicts whether complaint led to monetary relief.
- **Accuracy 60 %, Precision 70 % (relief class).**
- Top positive tokens: “refund,” “credited,” “charged back.”
- Top negative tokens: “dispute,” “investigation,” “pending.”
- Demonstrates potential to extract *explainable outcome signals* from free text.
- Language around refunds and reversals signals higher chance of monetary relief.

Impacts & Recommendations



At current thresholds, ~17% of complaints would be routed automatically, saving analysts ~X hours/week if each case takes ~5 min



Reduces manual routing effort, accelerates escalation.



Pilot recommendation: ATO + Zelle first, CNP next after dictionary refresh.



Maintain dashboard of topic spikes + new flagged narratives.

Limitations & Future Work



Weak labels: low recall, periodic dictionary refresh needed.



U.S. consumer focus → limited global coverage.



Future: add human-QA set (300–500 narratives), fine-tune DistilBERT, and monitor label drift.



This framework demonstrates that narrative-based triage can scale with minimal friction and strong privacy guarantees.