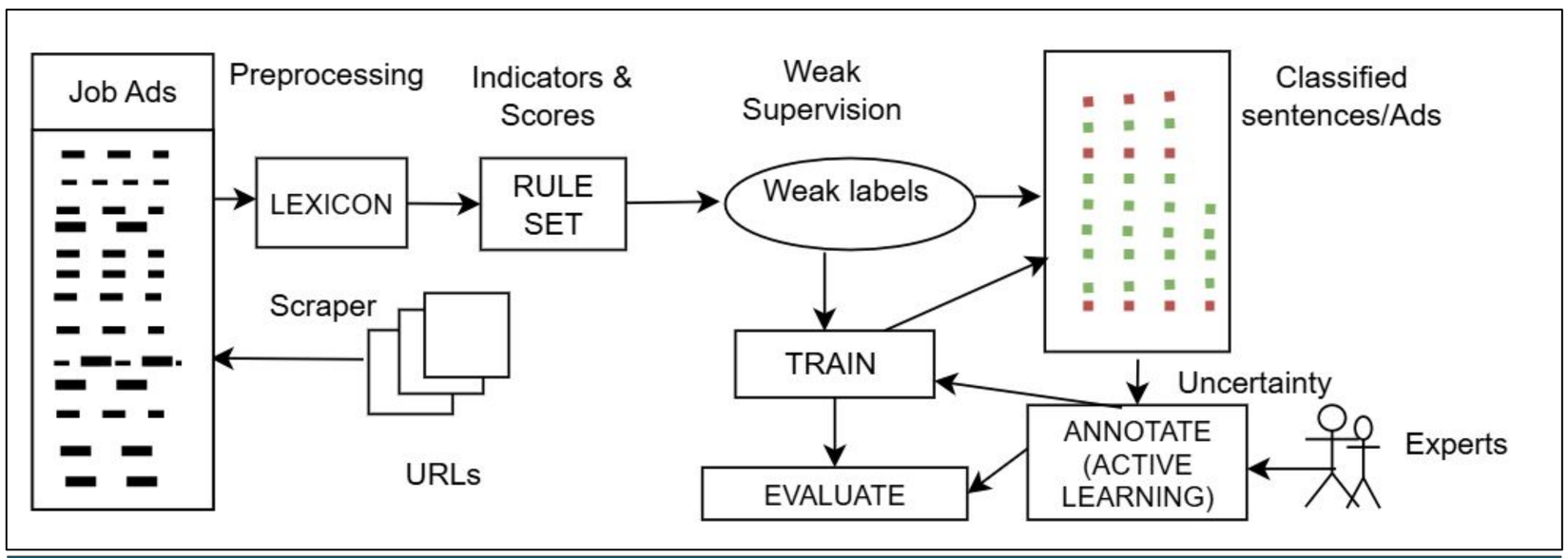
EXPLOIT Scan: A Domain-Specific, Label-Efficient Framework for Detecting Internet-Facilitated Labor Trafficking

Rebaka Pradhan, Onn Shehory

Information Systems, School of Business and Administration, Bar -Ilan University



Label-Efficient Detection Pipeline

Background

Traditional detection approaches rely on fully labeled datasets, but with trafficking datasets -

- labeling is expensive and expertise-dependent
- extreme class-imbalance.

In research:

- Recruitment through deceptive job advertisements (50%) (*UNODC, 2020)*
- Underexplored and underrepresented *(Sweileh, 2018) (Cockbain et al., 2018)*
- Need of empirical testing and operationalizing indicators for online space *(Volodko et al.,2020)*

AMP Model of Human Trafficking

Actions	Means	Purpose
Induce Recruits Harbors Transports Provides Or Maintains	Force Fraud or Coercion	Commercial Sex (Sex Trafficking) Or Labor/Services (Labor Trafficking)

Modeling recruitment deception + fraudulent means →predictive features of labor-exploitation risk.

A Job Ad (adapted from Workabroad.org)

ENGLAND – Warehouse Jobs! [SUPERMARKET BRAND] 250 vacancies
GUARANTEED EMPLOYMENT – IMMEDIATE START!
We are URGENTLY HIRING for logistics warehouses and offering permanent, legal and well paid jobs in factories, warehouses, printing houses and hotels in Great Britain.
We are one of the oldest well known agencies in Lithuania.
Up to 55–60 HOURS PER WEEK with weekly earnings of 320–500 GBP. WEEKLY PAY!

NO ENGLISH REQUIRED! NO EXPERIENCE NEEDED!

ACCOMMODATION PROVIDED and TRANSPORT ARRANGED.

We MEET YOU ON ARRIVAL and handle ALL PAPERWORK.

LIMITED SPOTS – DEPARTURES EVERY WEEK!

Jobs open for MEN and WOMEN. GOOD CONDITIONS and STABLE WORK.
Contact us via WEBSITE or PHONE for FREE CONSULTATION.

Research Q: Can an automated detection system effectively identify deceptive features that suggest internet- facilitated labor trafficking (IF-LT)?

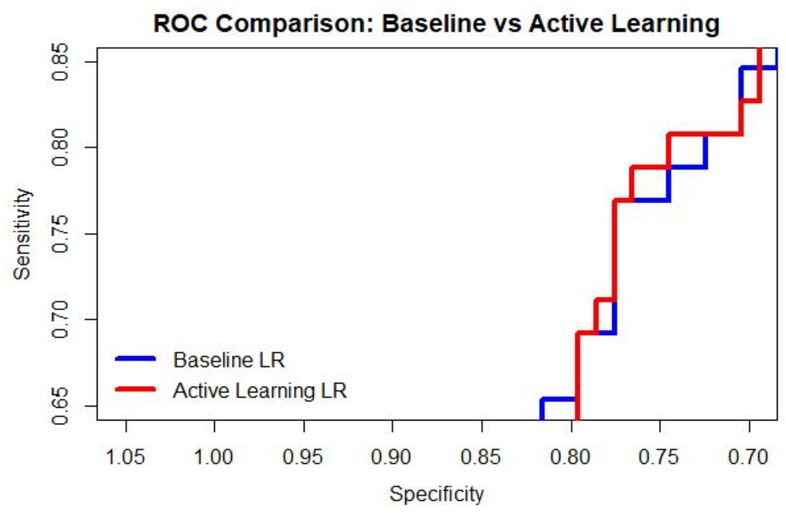
What we do (label -efficient modeling)

Pseudo Code

```
1: Input:
       D = ads, L = lexicon (DELPHI), E =
edge-case rules,n = gold set size, K = AL
batch size

2: Preprocess ads (clean → detect language →
translate)

3: Weak labeling for each ad x: # Lexicon
example
excessive_working_hours=
["bonus","shift","longer",
"flexible hours","lot of hours"]

score = Σ L matches
      # edge-case example
      if ("accommodation" & "transport" &
"insurance"): score += 4
      y□ = I[score > τ₀]

4: Sample n ads → gold labels
5: Train baseline f₀ on weak labels
6: Tune threshold τ on gold set

7: Active Learning:
      p = f(x); u = |p-τ|
      pick K most-uncertain → annotate →
retrain f

8: Output f* and τ*
```

Results (in -development)

Model performance	Threshold =0.21			
Model	**Accuracy**	**Recall**	**Specificity**	**F1**
Lexicon (weak labeling)	0.70	0.31	0.91	0.42
Logistic baseline	0.76	0.69	0.79	0.66
Logistic + AL round 1 (50 annotations)	0.77	0.75	0.77	0.69



Sensitivity is important to trafficking detection

Contribution

- **Novel:** first in–development framework to detect IF-LT in job ads.
- **Indicator driven:** UN/ILO indicator strengths → features.
- **Label- efficient:** low-cost, catches exploitative ads missed by single-indicator checks.
- **Real-world application:** potential screening tool for law enforcement

Next Steps

- Operationalize the wage/hour (exploitation indicators) information to national standard
  ad_wage < legal_minimum
  ad_hours > legal_max
- Active learning round 2 - crowdsourcing labels
- Further generalizability in a multilingual context
- Incorporating large language models (LLMs)

References

- Volodko, A., Cockbain, E., & Kleinberg, B. (2020). *Spotting the signs of trafficking recruitment online.* Trends in Organized Crime.
- UNODC (2018). *Human Trafficking Indicators.* https://www.unodc.org/pdf/HT_indicators_E_LOWRES.pdf
- ILO (2009). *Global Report on Forced Labour.* https://www.ilo.org/wcms_105023.pdf
- Ramchandani, P., Bastani, H., & Wyatt, E. (2025). *Unmasking Human Trafficking Risk in Commercial Sex Supply Chains with ML.* Manufacturing & Service Operations Management.
- Liu, J., Yu, H., Sujaya, V., Nair, P., Pelrine, K., & Rabbany, R. (2023). *SWEET: Weakly Supervised Person Name Extraction for Fighting HT.* Findings of EMNLP 2023.