

# On the Challenges of Creating Datasets for Analyzing Commercial Sex Advertisements to Assess Human Trafficking Risk and Organized Activity

Pablo Rivas,<sup>1</sup> Tomas Cerny,<sup>2</sup> Alejandro Rodriguez Perez,<sup>1</sup> Javier S Turek,<sup>3</sup> Laurie Giddens,<sup>4\*</sup> Gisela Bichler,<sup>5\*</sup> Stacie Petter<sup>6\*</sup>

\*Equal Contribution. <sup>1</sup>Department of Computer Science, School of Engineering & Computer Science, Baylor University. <sup>2</sup>Department of Systems & Industrial Engineering, College of Engineering, The University of Arizona. <sup>3</sup>Intelligent Systems Research, Intel Labs. <sup>4</sup>Information Technology & Decisions Sciences Department, University of North Texas. <sup>5</sup>School of Criminology & Criminal Justice, California State University, San Bernardino. <sup>6</sup>School of Business, Wake Forest University.

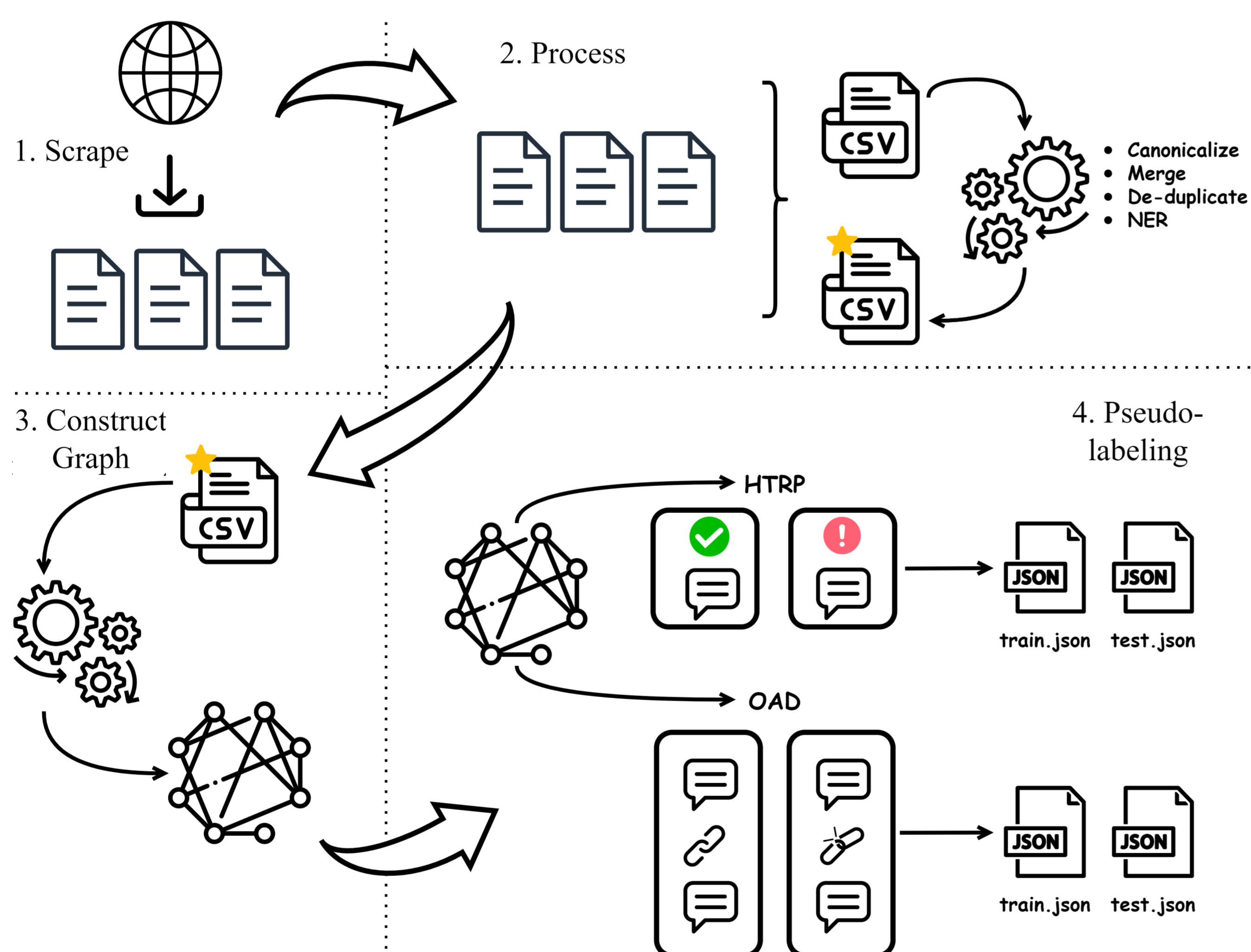
## Motivating Problem

- ▶ Securing **data for human trafficking research** is **notoriously challenging** and requires coordinated efforts.
- ▶ Criminals **disguise** their activities through coded messages, complicating research due to **evolving language**.
- ▶ A **reproducible approach** for collecting, processing, and labeling data from commercial sex ads is essential.

## Contributions

- ▶ Developed a reproducible and automated methodology to analyze five million commercial sex advertisements.
- ▶ Implemented custom-trained Named Entity Recognition (NER) models to handle adversarial text in ads.
- ▶ Constructed a Relatedness Graph to identify connections indicative of organized crime activities.
- ▶ Introduced pseudo-labeling techniques for Human Trafficking Risk Prediction & Organized Activity Detection.

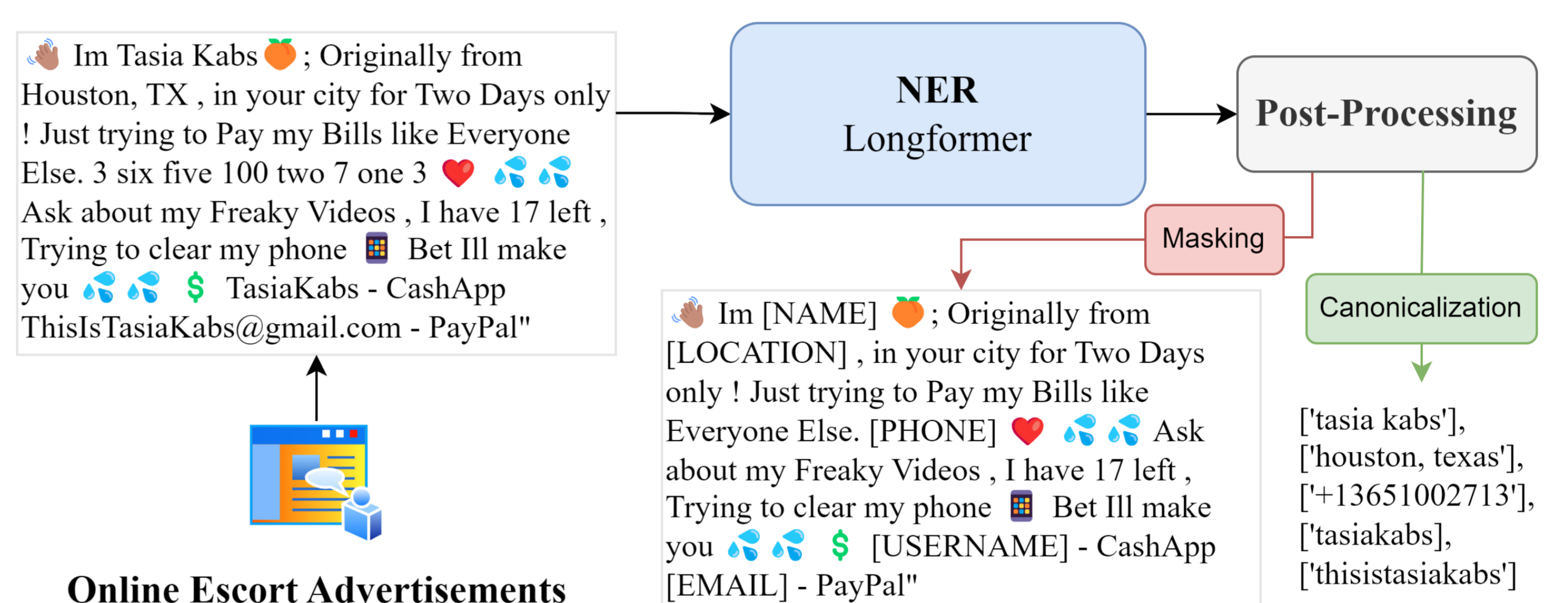
## Methodology Overview



We propose a methodology for identifying Human Trafficking Risk Prediction (HTRP) and Organized Activity Detection (OAD) from commercial sex ads, involving:



- Scrape:** Collect ads and metadata from a commercial sex website, addressing data cleaning and normalization.
- Process:** Apply deduplication techniques to refine the dataset, ensuring vital contextual cues are preserved.
- Graph:** Construct a Relatedness Graph to visualize connections between ads using shared identifiers.
- Pseudo-labeling:** Classify ad pairs and individual ads into binary categories for HTRP and OAD.

- ▶ At the end of the *Scrape* step, we end up with 5,053,249 ads, but after applying deduplication techniques, the dataset was of 515,865 unique ads, a **duplication rate of about 90%**.
- ▶ During the *Process* step we trained a NER model with expert ground-trut labeled data to address the limitations of standard NER models in **handling adversarial text**.



Online Escort Advertisements

- ▶ Connecting ads in the *Graph* step, using identifiers like phone numbers or social media handles, reveals **patterns of organized activity**, with connected components typical of **organized crime**, useful in the *pseudo-labeling* step.

Post 1  Warning: Graphic Content.   
NO MEETUPS ADD ME TO BUY VIDEOS, VIDEO CHAT OR SEXTING ONLYNew content of me getting fucked giving head taking cumshots facials anal creampie and girl on girl contact me or add me on Snapchat to purchase @[SNAPCHAT] serious inquiries only, I accept cashapp PayPal google pay Zelle chime or Facebook pay

Post 2  
No meetups add me or contact me to buy vids, sexting or video chat. Snapchat:[SNAPCHAT] Text [PHONE] Only add me if you're buying videos, video chat, sexting or custom videos I accept Cashapp Chime Apple Pay Zelle or Google pay

Two similar posts with a shared identifier.

Graph of connected components.

Size range	Components
1 node	184,877
2-10 nodes	51,117
10-100 nodes	5,928
100-1000 nodes	80
1000+ nodes	1
<b>Total</b>	<b>287,192</b>

Sizes of connected components.

A connected component labeled high risk by phone #.

## Summary of Challenges

- ▶ The heterogeneous nature of the data requires nuanced extraction and interpretation to preserve critical information.
- ▶ The presence of ad duplicates requires careful de-duplication to preserve meaningful content and unconventional word-number combinations and emojis.
- ▶ Selecting an effective NER tokenizer is pivotal, with Longformer emerging as the top choice for diverse data.
- ▶ A Relatedness Graph, although sparse, is essential for revealing patterns & facilitating the pseudo-labeling process.

Work partiall funded by the NSF under grants CNS-2136961 and CNS-2210091.



NAACL 2024

