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

Pablo Rivas[1*] **Tomas Cerny**[2] **Alejandro Rodriguez Perez**[1]
Javier S Turek[3] **Laurie Giddens**[4] **Gisela Bichler**[5] **Stacie Petter**[6]
[1]Baylor University [2]The University of Arizona [3]Intel Labs [4] University of North Texas
[5]California State University San Bernardino [6] Wake Forest University
[*]Pablo_Rivas@Baylor.edu

Abstract

Our study addresses the challenges of building datasets to understand the risks associated with organized activities and human trafficking through commercial sex advertisements. These challenges include data scarcity, rapid obsolescence, and privacy concerns. Traditional approaches, which are not automated and are difficult to reproduce, fall short in addressing these issues. We have developed a reproducible and automated methodology to analyze five million advertisements. In the process, we identified further challenges in dataset creation within this sensitive domain. This paper presents a streamlined methodology to assist researchers in constructing effective datasets for combating organized crime, allowing them to focus on advancing detection technologies.

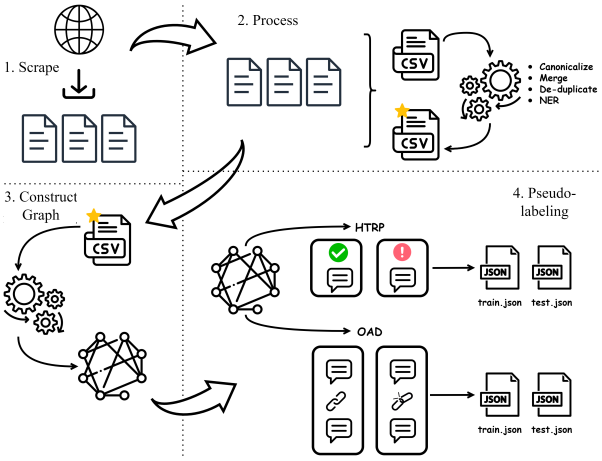


Figure 1: Methodology to generate a pseudo-labeled dataset in human trafficking risk prediction and organized activity detection tasks.

1 Introduction

The landscape of commercial sex advertisements is not just a platform for services but also a lucrative target for human traffickers and organized crime to exploit for financial gain. For law enforcement, the challenge is monumental – the volume and ever-renewing stream of ads make it almost impossible to keep up (Giddens et al., 2023). With the evolution of NLP, there is a promising path forward to aid in identifying these suspicious ads; however, current approaches hinge on the availability and reliability of the datasets, lacking automation and reproducibility (Vajiac et al., 2023).

In our pursuit to bolster the efforts of criminal investigators in detecting illicit activities, we embarked on a two-year journey to compile a comprehensive dataset of commercial sex ads, using the methodology depicted in Figure 1. This endeavor initially appeared straightforward and unveiled various unexpected hurdles and obstacles. We intend not to showcase the dataset but to share the lessons learned when creating it. Due to the rapid pace at which this data can become obsolete, we focus on the methodology behind building this dataset. We aim to illuminate the challenges and pitfalls encountered along the way, guiding fellow researchers. This insight will enable others to sidestep these challenges and more efficiently contribute to the collective fight against online crime.

2 Paucity of Datasets in This Domain

Securing data for human trafficking (HT) research is notoriously challenging, requiring the coordinated efforts of academics, law enforcement, and, at times, victims themselves. This collaboration is complex and hard to manage. Innovative approaches, including using heuristics as stand-ins for direct indicators of trafficking and semi-automated labeling, have helped researchers sidestep these issues. Yet, accessing even the most basic raw data has proven difficult, demanding significant effort to extract from public sources, often with limited outcomes (Dubrawski et al., 2015; Portnoff et al., 2017; Hundman et al., 2018).

Language is a crucial tool in online illicit activities, including HT. Criminals disguise their activi-

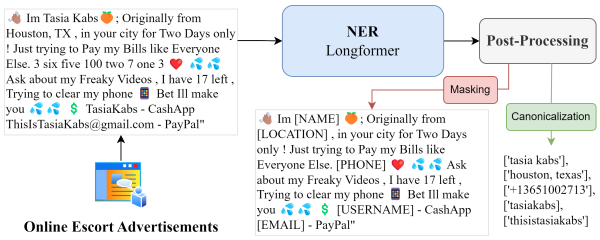


Figure 2: Processing the text of an ad with the NER pipeline. *Personally identifiable data has been changed.*

Table 1: Size of the 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
Total	**287,192**

ties and intentions through coded messages, crafting a specialized language that evolves with cultural and societal shifts. This evolution complicates research, as traffickers continually alter their communication to evade detection, making any model based on static data quickly outdated (Dubrawski et al., 2015; Latonero, 2011). Presenting a solid, reproducible approach for collecting, processing, and labeling data from commercial sex ads will significantly bolster the creation of current datasets.

3 Methodology Overview

Embarking on a project to identify Human Trafficking Risk Prediction (HTRP) and Organized Activity Detection (OAD) from commercial sex advertisements, we devised a methodology demanding minimal human oversight. Illustrated in Figure 1, our approach navigates through data scraping, processing, and analysis, leading to a graph-based model to identify risk of trafficking and organized crime.

Initially, we scrape a website for commercial sex to collect ads and metadata, uniformizing data. This process involves data cleaning and normalization, addressing the challenge of ensuring that vital contextual cues like slang or emojis are not lost, contrary to conventional text processing practices (Zhu, 2019; Wiriyakun and Kurutach, 2022). To manage data diversity and address the issue of data duplication—where our analysis found a staggering 90% rate of textual duplication—we applied deduplication techniques, refining our dataset to 515,865 unique ads out of an original total of 5,053,249 ads (Rodriguez and Rivas, 2023).

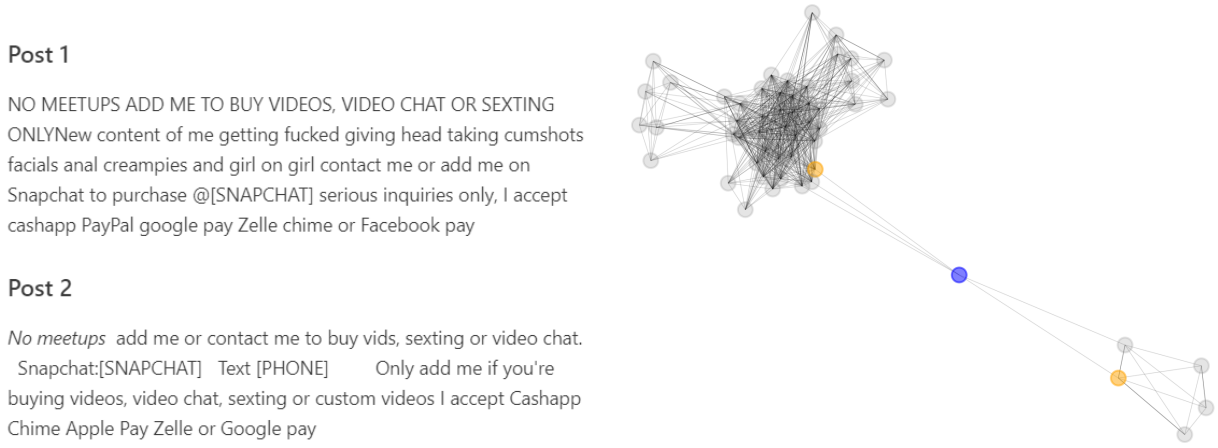
For Named Entity Recognition (NER), illustrated in Figure 2, we encountered limitations with standard models. These models faltered against the adversarial text common in such ads, prompting us to custom-train NER models on a dataset of 1,810 labeled ads, identifying entities relevant to our investigation. Our exploration of various NER models highlighted the superiority of Longformer and XLNet, with Longformer slightly edging out due to its handling of out-of-vocabulary tokens.

Constructing a Relatedness Graph from the deduplicated posts allowed us to visualize connections between ads through shared identifiers, like phone numbers or social media handles. Despite the vast potential connections, the graph revealed a sparse structure dominated by isolated nodes and significant connected components indicative of organized activities, as shown in Figure 3 and in Table 1.

Pseudo-labeling for OAD and HTRP involved leveraging the Relatedness Graph to classify ad pairs and individual ads, respectively, into binary categories. The division of this graph into connected components and their subsequent distribution into training and test sets underscored the complexities inherent in dataset preparation, especially given the unbalanced nature of real-world data. Our methodical approach to OAD involved the binary labeling of ad pairs based on their interconnectedness, emphasizing the importance of a balanced dataset and the necessity to discard excessively similar advertisements to mitigate bias. This was operationalized through a similarity threshold, informed by the Levenshtein distance, set at 0.5 after rigorous evaluation. For HTRP, we ventured beyond mere text analysis, employing heuristics based on ad metadata—such as the distance between locations exceeding 300 miles and the number of unique identifiers—to infer trafficking risk.

4 Limitations

Our research encountered a few important limitations. Sparse data, NER pipeline errors, and data source heterogeneity introduce significant variability and potential inaccuracies. The process's susceptibility to bias—evidenced by differences in labeling driven by the algorithm's reliance on hard identifiers and geographical heuristics—raises concerns, as confirmed by statistical testing (Wilcoxon, 1945); more specifically, a Wilcoxon rank-signed



(a) (b)

Figure 3: Ads connected indirectly in a connected component. (a) Description text of the highlighted posts. *Personally identifiable data has been changed.* (b) Connected component graph where referred posts are in orange.

test with a p-value of 0.004. These limitations, underscored by the variance in data treatment and the subjective selection of similarity metrics, highlight the complexities of deploying NLP techniques in the sensitive context of human trafficking, suggesting that while our approach marks a step forward, it navigates a landscape riddled with challenges.

5 Discussion of Challenges

In constructing our dataset, we encountered critical challenges that underscore the complexity of this endeavor. First, the data's heterogeneous nature often meant that while some fields were missing, crucial information could still be embedded within text or images, necessitating a nuanced approach to extraction and interpretation. The presence of ad duplicates called for careful de-duplication, emphasizing the importance of thoughtful preprocessing to preserve meaningful content, such as unconventional word-number combinations or emojis, which might otherwise be overlooked. Selecting an effective NER tokenizer was pivotal; *Longformer* emerged as our top choice, adept at handling the diverse data we encountered, including extracting and consolidating obvious metadata from ad descriptions. Post-processing steps, including normalizing entities like phone numbers for edge formation in the Relatedness Graph, were essential for creating meaningful connections between data points. This graph, although sparse, served as a critical backbone for our analysis, revealing interesting patterns and insights, as evidenced by Figure 4 and highlighting the skewed distribution of component sizes in our dataset (Table 1). Ultimately, the

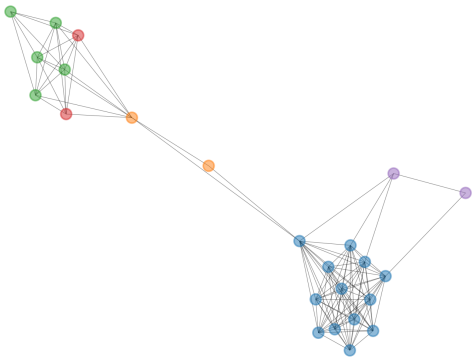


Figure 4: A connected component labeled positive due to several phone numbers found. Each color represents a different phone number encountered.

Relatedness Graph facilitated the pseudo-labeling process for OAD and HTRP, showcasing our research's intricate interplay of challenges.

6 Conclusion

Our research presents a methodology centered around the complexities of detecting organized crime, particularly in human trafficking. By integrating advanced NER techniques and utilizing the Longformer model for its adept handling of extensive texts, we have developed a dynamic dataset creation process that identifies non-trivial connections within ads. Our work emphasizes the importance of adaptable, privacy-aware approaches in dataset development, offering the research community a refined, scalable framework for navigating the initial challenges of data-centric investigations into organized crime.

Acknowledgements

This work was funded by the National Science Foundation under grant CNS-2210091.

References

Artur Dubrawski, Kyle Miller, Matthew Barnes, Benedikt Boecking, and Emily Kennedy. 2015. Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking*, 1(1):65–85.

Laurie Giddens, Stacie Petter, Gisela Bichler, Pablo Rivas, Michael H. Fullilove, and Tomas Cerny. 2023. Navigating an interdisciplinary approach to cybercrime research.

Kyle Hundman, Thamme Gowda, Mayank Kejriwal, and Benedikt Boecking. 2018. Always lurking: Understanding and mitigating bias in online human trafficking detection. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 137–143, New York, NY, USA. Association for Computing Machinery.

Mark Latonero. 2011. Human trafficking online: The role of social networking sites and online classifieds. *Available at SSRN 2045851*.

Rebecca S. Portnoff, Danny Yuxing Huang, Periwinkle Doerfler, Sadia Afroz, and Damon McCoy. 2017. Backpage and bitcoin: Uncovering human traffickers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1595–1604, New York, NY, USA. Association for Computing Machinery.

Alejandro Rodriguez and Pablo Rivas. 2023. Combatting human trafficking in the cyberspace: A natural language processing-based methodology to analyze the language in online advertisements.

Catalina Vajiac, Meng-Chieh Lee, Aayushi Kulshrestha, Sacha Levy, Namyong Park, Andreas Olligschlaeger, Cara Jones, Reihaneh Rabbany, and Christos Faloutsos. 2023. Deltashield: Information theory for human- trafficking detection. *ACM Trans. Knowl. Discov. Data*, 17(2).

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Chawit Wiriyakun and Werasak Kurutach. 2022. Extracting co-occurrences of emojis and words as important features for human trafficking detection models. *Journal of Intelligent Informatics and Smart Technology. 2021a*, 7.

Jessica H Zhu. 2019. *Detecting food safety risks and human tracking using interpretable machine learning methods*. Ph.D. thesis, Massachusetts Institute of Technology.

A Ethics and Broader Impact Statement

This research addresses the controversial and sensitive issue of detecting human trafficking within online commercial sex advertisements. Our primary goal is to identify linguistic traits that can help understand criminal communication in consumer-to-consumer online marketplaces.

To protect potential victims of trafficking, we have chosen not to release the dataset. Instead, we provide a detailed protocol to allow reproducibility without compromising safety and privacy. This ensures that sensitive data is not exposed, minimizing the risk of harm to vulnerable individuals.

We acknowledge the potential misuse of our research, which could inadvertently target legitimate sex workers. To mitigate this risk, our findings only highlight patterns and indicators. It is crucial that any findings derived from our methodology be used as part of a broader, victim-centered approach prioritizing safety and well-being over punitive measures.

Our ethical considerations include:

- **Victim Protection:** By withholding the dataset and focusing on methodological transparency, we prevent potential harm from the misuse of sensitive data.
- **Responsible Data Use:** We urge researchers and practitioners to collaborate with social scientists, legal experts, and victim advocacy groups to ensure ethical use of our protocol.
- **Contextual Analysis:** Our methodology should be used as a supplementary tool within a holistic investigative framework that includes qualitative assessments and corroborative evidence.
- **Stakeholder Impact:** We recognize the sensitive nature of our research and its potential impact on various stakeholders, including law enforcement, policymakers, researchers, and victims. Our goal is to contribute positively to combating human trafficking.

By prioritizing victim protection, promoting responsible data use, and encouraging a holistic approach to interpretation, we aim to make a meaningful contribution to the fight against human trafficking. Our commitment is to ensure that our work is used ethically and effectively to aid in identifying and protecting victims, while preventing misuse that could cause harm.