Unmasking human trafficking risk in commercial sex supply chains with machine learning

Pia Ramchandani
Operations, Information and Decisions, Wharton School, University of Pennsylvania, piar2@wharton.upenn.edu

Hamsa Bastani
Operations, Information and Decisions, Wharton School, University of Pennsylvania, hamsab@wharton.upenn.edu

Emily Wyatt
Uncharted Software, Founding Partner of the TellFinder Alliance, ewyatt@tellfinder.com

Problem Definition: The covert nature of sex trafficking provides a significant barrier to generating large-scale, data-driven insights to inform law enforcement, policy and social work. Existing research has focused on analyzing commercial sex sales on the internet to capture scalable geographical proxies for trafficking. However, ads selling commercial sex do not reveal information about worker consent. Therefore, it is challenging to identify risk for trafficking, which involves fraud, coercion or abuse.

Methodology: We leverage massive deep web data (collected globally from leading commercial sex websites) in tandem with a novel machine learning framework (combining natural language processing, active learning and network analysis) to study how and where sex worker *recruitment* occurs. This allows us to unmask deceptive recruitment patterns (e.g., an entity that recruits for modeling, but sells sex). Our analysis provides a geographical network view of commercial sex supply chains, highlighting deceptive recruitment-to-sales pathways that signal high trafficking risk.

Managerial Implications: Our results can help law enforcement agencies along trafficking routes better coordinate efforts to tackle trafficking entities at both ends of the supply chain, as well as target local social policies and interventions towards exploitative recruitment behavior frequently exhibited in that region.

Key words: human trafficking, machine learning, deep web, active learning, networks, text analytics

1. Introduction

According to the FBI, sex trafficking is the fastest growing organized crime business and the third largest criminal enterprise in the world (Walker-Rodriguez et al. 2011). The International Labor Organization estimates there were 4.8 million sex trafficking victims in 2017 alone (ILO 2017). Consequently, there is high demand from field experts (Office 2006, Laczko 2002, Witte 2018) and academics (Flynn et al. 2014, Androff 2011, Orme and Ross-Sheriff 2015, Kotrla 2010, Potocky 2010) for a large-scale and data-driven view of the underlying supply chain dynamics (Roby and Vincent 2017) of trafficking that can inform law enforcement, policy and social work. For instance, understanding where and how victims are recruited in different regions can enable preventative interventions at the source of the supply chain (recruitment) in contrast to prevalent mitigation

strategies that target the end of the supply chain (sales) (Shively et al. 2012, Murphy 2016). Furthermore, inferring likely recruitment-to-sales trafficking routes of criminal entities can enhance coordination strategies between relevant law enforcement agencies and task forces to increase efficiency of counter-trafficking efforts (Heilemann and Santhiveeran 2011, Hodge and Lietz 2007, Johnson 2012, Jones et al. 2007, Roby 2005).

However, the covert nature of trafficking provides a significant barrier to generating such insights. For example, limited existing research literature on sex trafficking uses any data, and those that do primarily leverage qualitative interviews with trafficking survivors (Okech et al. 2018). It is hard to generate quantitative and generalizable insights from such interviews, because they are qualitative in nature and severely limited in scale; moreover, they can be traumatic for victims and can result in unreliable information (Androff 2011).

In this paper, we use unstructured, massive deep web data to characterize trafficking recruitment and sales risk at scale. The deep web represents portions of the World Wide Web that are not indexed by traditional search engines, e.g., temporary or dynamic content from private websites that can only be accessed via specialized queries. A significant portion of commercial sex activity – and the exploitative behavior that accompanies it – occurs online (Raets and Janssens 2019, Latonero 2011), making the deep web a rich and relevant data source. Trafficking is commonly targeted at vulnerable populations (e.g., 1 out of 5 homeless youth in top cities in the United States and Canada have been identified as victims of human trafficking (Murphy 2016)), who are frequently recruited online through "fishing" strategies that offer well-paid jobs to attract potential victims to make initial contact with traffickers (Kangaspunta et al. 2020).

We begin by leveraging data from leading commercial sex advertisement websites in conjunction with a novel machine learning framework to construct a geographical network view of commercial sex *supply chains*, from recruitment to sales. Existing research has focused solely on analyzing commercial sex sales to capture proxies for trafficking (Dubrawski et al. 2015, Zhu et al. 2019). Importantly, however, commercial sex and sex trafficking are not synonymous (Albright and D'Adamo 2017): "Unlawful commercial sex acts overlap with sex trafficking when participation occurs by means of force, fraud, or coercion..." (Dank et al. 2014). In other words, it is critical to understand how victims are recruited into the commercial sex supply chain to distinguish trafficking victims and commercial sex workers.

To address this challenge, we study how and where *recruitment* occurs in the supply chains we uncover. Indeed, in over 50% of trafficking cases analyzed by the UN Office of Drugs and Crime (UNODC) in 2020, victims reported making initial contact with a trafficker in response to a deceptive job advertisement. For example, in one case, traffickers recruited approximately 100 women through a modeling job posting and then sex trafficked these women (Kangaspunta

et al. 2020). Traffickers have also recruited for other adult services (e.g., stripping) before forcing victims into sex sales (UNODC 2020). Thus, if an entity recruits victims through non-sex offers (e.g., purportedly for modeling or massage) and is also involved in commercial sex sales, then this is an informative indicator that trafficking may have occurred. This proxy was informed through close collaboration with domain experts from the Tellfinder Alliance, ranging from members of human trafficking taskforces to policymakers. However, it is important to note that not all instances identified in this manner necessarily correspond to human trafficking; rather this is a high-quality proxy. Our work captures trafficking *risk* by linking deceptive (non-sex) recruitment offers to commercial sex sales by the same entities.

However, identifying recruitment content in ads has historically been a significant hurdle due to the nature of sex trafficking: while sex sales ads are prevalent and convey clear intent to consumers, recruitment ads are sparse and are typically designed to trick potential victims into being trafficked. Thus, while recent work has developed techniques to scrape deep web data, extract relevant meta data (e.g., phone numbers, email addresses) and convert it into databases that support trafficking investigation inquiries by law enforcement agencies (TellFinder 2021, Kejriwal and Kapoor 2019, Zhang et al. 2017), such data has not been used for large-scale analysis of commercial sex supply chains, primarily due to the difficulty in identifying recruitment from unstructured text. We address this challenge through a novel machine learning framework that combines natural language processing, active learning, network properties, and domain expertise to distinguish recruitment and sales content at scale. We then leverage shared meta data to infer trafficking risk in commercial sex supply chain networks.

Our results yield substantial insights into the structure of commercial sex supply chains, including several policy-relevant insights. First, while sex sales predominantly occur in large urban centers, we find evidence that recruitment is concentrated in suburban, economically constrained areas. Furthermore, there is significant variation in how vulnerable populations are recruited in different locations, suggesting opportunities for targeted job search training (Murphy 2016). By highlighting links between deceptive (non-sex) recruitment offers and sex sales made by the same entity, we are uniquely able to infer likely trafficking routes between cities. Importantly, these routes can help inform coordination strategies between relevant law enforcement agencies.

1.1. Additional Related Literature

The field of operations management is uniquely positioned to help tackle challenges in countering sex trafficking (Konrad et al. 2020). As described above, we take a supply chain perspective, aiming to understand how sex workers are recruited (supply) and sold (demand) by trafficking entities. A few studies have applied operations techniques to empirically and theoretically analyze

other challenges in sex trafficking (Konrad et al. 2017). One important problem is effectively allocating resources for social policy interventions; Kaya et al. (2022) and Chan et al. (2018) use a multidimensional knapsack algorithm to improve access to housing and support services for homeless youth, with the goal of mitigating their vulnerability to trafficking. Maass et al. (2020) optimizes the placement of shelters for human trafficking survivors to maximize societal welfare while respecting budgetary constraints. More related to our work, another stream of work aims to support law enforcement by signaling businesses or entities that are likely to be involved in trafficking. For example, Li et al. (2021) uses online customer review data to identify massage businesses that are likely selling sex. Keskin et al. (2021) examines and predicts the movement patterns of entities selling commercial sex; this is an important goal to ensure that law enforcement officials do not pursue "wasted" stings on entities that have already left a location. Kosmas et al. (2020) model interdiction strategies that can disrupt these illicit networks (Kosmas et al. 2020). Unlike past work which has focused on sex sales, our work identifies deceptive recruitment of victims, thereby providing an informative signal for distinguishing commercial sex and human trafficking.

2. Deep Web Data

Our core deep web dataset is obtained from our collaborators at the TellFinder Alliance for global counter-human trafficking (TellFinder 2021). The deep web consists of (often temporary) pages that are not indexed by Google, and therefore need to be scraped in real-time. TellFinder works with its partners in law enforcement to identify websites with significant commercial sex activity – which often carry risk of exploitation and human trafficking (Kangaspunta et al. 2020) – that are relevant to counter-trafficking efforts. They leverage recent technology developed to scrape deep web data, extract relevant meta data (e.g., phone numbers, email addresses) and convert it into databases that support trafficking investigation inquiries by law enforcement agencies (TellFinder 2021, Hall et al. 2015).

There are several kinds of websites where commercial sex activity can be deduced. These include service review websites such as the Erotic Review or Rubmaps and discussion forms (where content is largely shared by consumers, rating specific sexual services), as well as commercial sex advertisement websites (where content is largely shared by entities selling sexual services). Since our primary goal is to identify supply chains of specific entities – i.e., connect deceptive recruitment offers to sex sales by the same entity in order to pinpoint human trafficking risk – we focus on commercial sex advertisement websites where we can extract identifying information (e.g., phone numbers, emails) of the entities selling sex. Table 1 shows summary statistics of different identifiers extracted from posts on our deep web dataset; indeed, we see that the large majority of posts contain identifying information that can be used to connect entity-specific activity.

Identifier	# Unique Occurrences	# Posts Including Identifier	% Post Including Identifier
Phone number	393,132	8,503,617	62.7%
Email	214,728	1,489,803	11.0%
Social media handle	8,645	395,547	2.92%
Username	44,007	44,235	0.33%
Location	1,364	12,716,641	93.7%

Table 1 Deep web data sample summary.

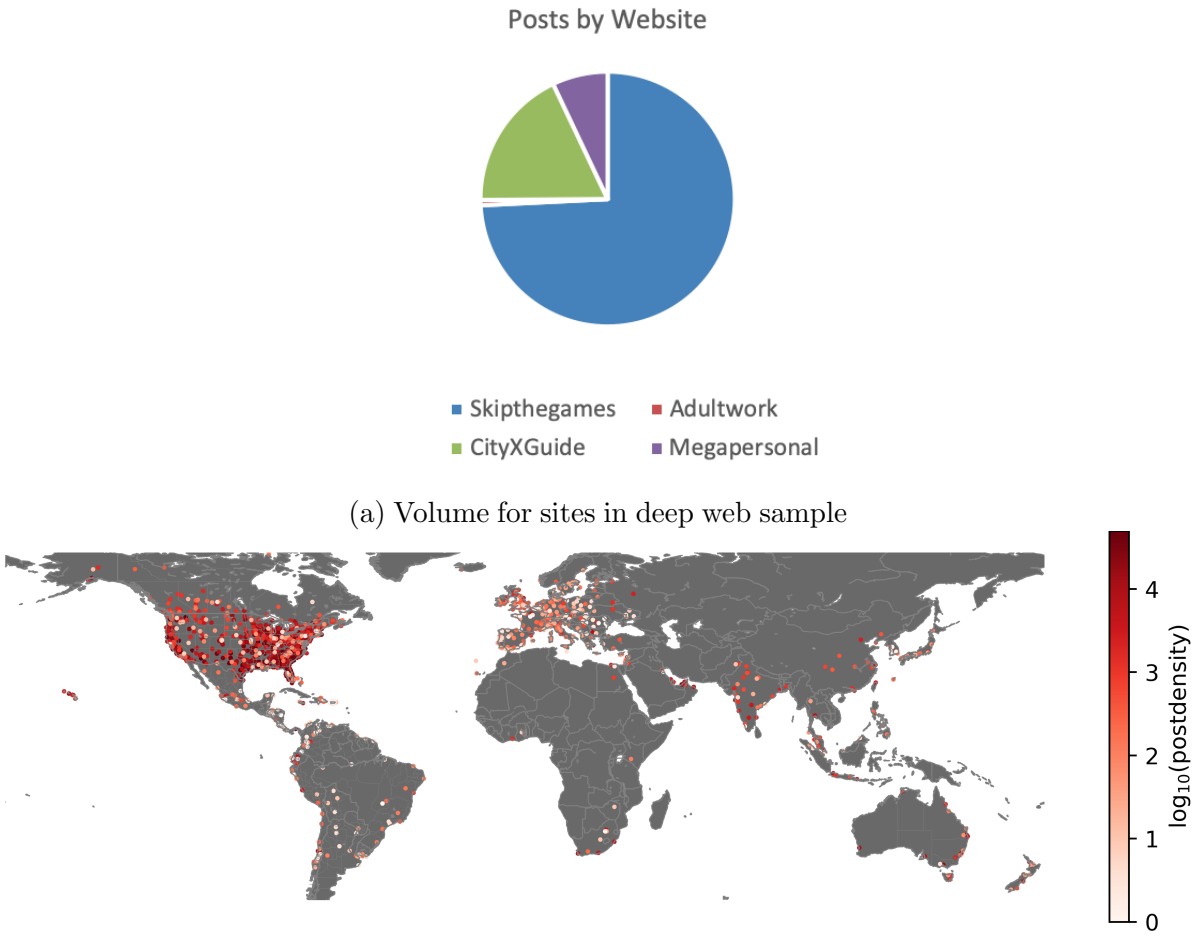
One may not a priori expect that significant recruitment activity occurs on websites that primarily advertise commercial sex. We learned of this behavior from our law enforcement partners (e.g., when running a phone number associated with a criminal case through the TellFinder tool, one partner found that the number was associated with both sales posts and deceptive recruitment offers on the same website, thereby providing supporting evidence that this was likely a human trafficking case). However, it was not possible to study these recruitment offers at scale since they are extremely rare; our machine learning framework addresses this challenge, and indeed identifies thousands of deceptive recruitment offers spanning many different tactics on these websites. These included job postings, personal ads, or ads offering other types of skills (see Appendix B). Separately, we also examined posts on Craigslist.com, SpaStaff.com, and Indeed.com, where general recruitment activity is common (e.g., in the jobs and services categories) and content is not focused on commercial sex sales. If identifiers extracted from commercial sex advertisement websites match identifiers for entities recruiting on the websites for non-sex employment, this can also suggest high risk for sex trafficking. However, we found only 4 such matches to Craigslist and no matches to SpaStaff.com or Indeed.com, suggesting that this behavior is relatively uncommon.

Our deep web dataset spans four websites that advertise commercial sex through English language posts. These posts were collected over a 9-month period spanning July 1, 2017 to September 6, 2018. The websites, in order of volume, include:

- www.skipthegames.com
- www.cityxguide.app
- www.megapersonals.eu
- www.adultwork.com

The resulting dataset comprises of 13,568,130 posts over 428 unique days. Figure 1 shows the breakdown of posts across website and location.

Figure 1 shows a key limitation of focusing on English language posts: the geographical distribution of our data is concentrated in countries with large English-speaking populations. In particular, approximately 95% of the posts are from the United States, Canada, the United Kingdom and Australia. We do find significant sales and recruitment activity in the rest of Europe and in India,



(a) Volume for sites in deep web sample

(b) Geographical distribution of deep web data sample posts

Figure 1 **Count of deep web posts in our dataset by (a) website, and (b) location.**

but this may be a biased sample since it omits activity occurring in local languages in those countries. A promising direction of future work is adapting our approach to other languages to improve global coverage. Note that this would require domain experts who speak these local languages to operationalize our human-in-the-loop and active learning steps.

3. Machine Learning Framework

Our analysis is based on data supplied by the TellFinder Alliance, a network of law enforcement, technology, research, and nonprofit partners focused on adapting analytics and deep web data for counter-human trafficking applications (TellFinder 2021). This dataset includes approximately 14 million public, English-language commercial sex advertisements (which we call posts) collected from adult services websites in the deep web over a 9-month period with global geographic coverage. Nearly all posts are related to commercial sex sales, while a small subset are related to recruitment for different types of jobs (e.g., modeling, massage, etc.). Moreover, each post is associated with

meta data (i.e., any extracted phone numbers, email addresses, social media handles, etc.), which can be used to connect pairs of posts that are made by the same entity.

Our first step is to infer the underlying commercial sex supply chain. To this end, we train a deep neural network that distinguishes recruitment from sales posts based on the unstructured text in that post. A priori, all posts are unlabeled. Labels must be obtained by having a domain expert manually read the content of each post and assign a label (recruitment vs. sales); recruitment posts are additionally categorized into types (e.g., sugar parent) based on the type of employment offer made. Manually labeling all 14 million posts is clearly infeasible; instead, we design an active learning approach to train a model with as few labels as possible. We face two challenges:

1. **Extreme Data Imbalance:** We estimate that only 0.06% of posts are recruitment-related, while the rest are sales, i.e., one would have to manually label nearly 2000 randomly chosen posts to find a single instance of recruitment in expectation. Thus, traditional supervised or active learning techniques, which rely on an initial well-balanced training set, are infeasible.

2. **Objective Mismatch:** We seek to identify different recruitment approaches across many locations. For instance, one auxiliary task is to identify pairs of posts (one recruitment and one sales) in different locations that are linked to the same entity by their meta data; such a pair corresponds to a potential edge in the supply chain network. Thus, traditional active learning techniques that focus purely on overall accuracy may be insufficient.

We leverage weak learners (Zhou 2018, Ratner et al. 2017a) in conjunction with active learning (Gonsior et al. 2020, Nashaat and Miller 2021) to address these challenges (see first two panels of Figure 2). We give an overview of our approach in what follows.

3.1. Initial Training Set

Our entire dataset of 13,568,130 posts is initially unlabeled. Labels must be obtained by a domain expert manually reading the content of a post and assigning a label (recruitment vs. sales); recruitment posts are additionally tagged with the type of recruiting tactic. The sensitive nature of the data (i.e., containing personal identifiable information such as names and contact information) precludes a crowd-sourcing approach. To the best of our knowledge, there has been no prior work in academia or industry on predicting trafficking risk in recruitment using machine learning. As a result, we cannot apply existing models to our unlabeled data. Thus, we must obtain manual labels for a subset of our posts in order to obtain an initial training set to build a predictive model.

A common approach is to label a random subset of the posts to create this initial training set (Olsson 2009). However, as noted in the main text, we estimate that over 99.4% of posts are aimed at sex sales; this is to be expected since our dataset is collected from leading commercial sex advertisement websites. Consequently, a domain expert would have to manually label nearly 2000

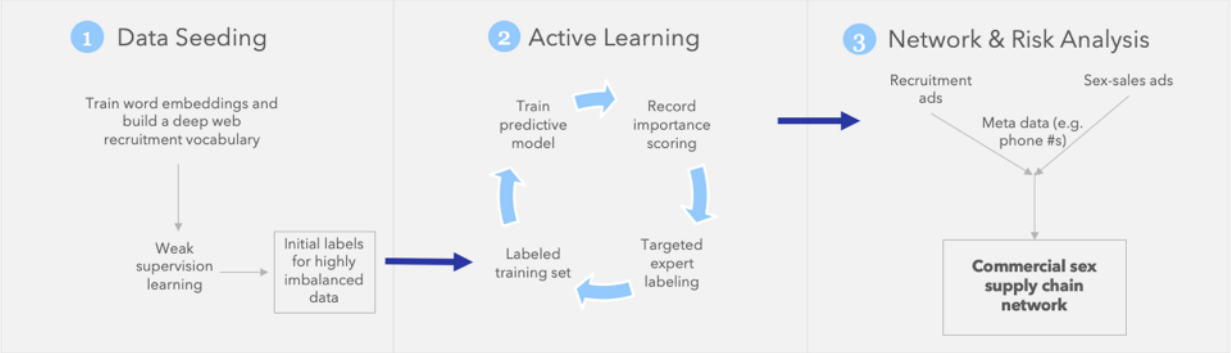


Figure 2 **Summary of our machine learning framework.We first train domain-specific word embeddings and collect expert-identified terms to develop a 'recruitment vocabulary.' This informs a weak learning heuristic to identify an initial well-balanced training set. We then apply active learning techniques (additionally incorporating geographical diversity and the likelihood of identifying new network connections) to iteratively label additional posts and update our predictive model until its performance converges. Finally, we connect recruitment and sales activity via meta data to identify supply chain networks.**

randomly chosen posts to find a single instance of recruitment with reasonable likelihood. More advanced sampling approaches – e.g., sampling from clusters of the data (Dligach and Palmer 2011) or dense regions (McCallumzy and Nigamy 1998), or maximizing diversity in the sample (Yang et al. 2015) – experience similar imbalance issues in constructing an initial training set. This is problematic because, from a statistical perspective in a classification problem, the effective sample size of the data scales with the number of observations in the minority class (i.e., the number of labeled recruitment posts). At the same time, we leverage deep learning models (due to their incomparable success on prediction with unstructured text (Minaee et al. 2021)), which have a high tendency to overfit training data and are therefore data hungry (Chen and Lin 2014).

Thus, we must carefully choose a subset of posts that (i) has a far higher likelihood of containing recruitment posts, thereby ensuring that our initial training set has a nontrivial effective sample size, and (ii) is of a manageable size for manual labeling by domain experts. To address this issue, we construct an initial 'recruitment vocabulary' that informs a human-in-the-loop weak learning approach.

First, we must preprocess the text to capture the semantic content of words in a way that can be passed as an input to a machine learning algorithm. A leading approach is to train word embeddings, which project words into a vector space whose distance metric captures semantic similarity (Mikolov et al. 2013). Typically, word embeddings are trained to encode how frequently pairs of words co-occur in text; this is an effective approach since words with similar meanings tend to occur in similar contexts. To capture the unique context of our data, we train our own domain-specific word embeddings using Gensim word2vec (Rehurek and Sojka 2010, Gensim 2021),

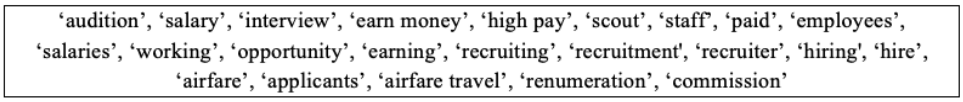


Figure 3 List of expert-identified recruitment terms used to inform weak learners.

which uses "continuous bag of words" (CBOW). Specifically, CBOW involves specifying a window of "context words" around a "target" word that are used to predict the target. Model parameters are iteratively updated using different pairs of context-target word combinations to modify a target word's embedding, based on its appearance with its co-occurring neighbors (Rong 2014). Following standard pre-processing techniques in natural language processing (Symeonidis et al. 2018), we drop stop words (e.g., 'the', 'is', 'but') and lemmatize the vocabulary (e.g., 'caring' would be converted to care, 'communicating' would be converted to communicate). This leaves a unique vocabulary of size 223,883 across all posts. Using a context window size of 5, we train embeddings of dimension 100.

Then, we identify some candidate terminology that signifies recruitment risk from discussions with domain experts. These words were chosen through a human-in-the-loop process to maximize the likelihood of the corresponding post being recruitment-related; thus, words such as 'model' that are likely to appear in both recruitment and sales posts were excluded in order to avoid a high false positive rate. Our initial 'recruitment vocabulary' includes all terms whose embeddings are within a short distance of the embeddings corresponding to the expert-identified terms shown in Figure 3.

As with traditional weak supervised learning (Ratner et al. 2017a), the presence of a term from our recruitment vocabulary provides a noisy signal that a post may be related to recruiting. Using the Snorkel package (Snorkel 2021), we train a weak supervision model that results in 1651 posts containing part of this recruitment vocabulary. We then obtain labels for this small subset of posts, resulting in 369 recruitment-related posts. Note that this corresponds to 22% of the labels being positive, compared to only 0.06% of the labels being positive on a random subsample of our dataset. However, this training dataset is biased based on the knowledge of domain experts and only uncovered 3 types of recruitment templates. Therefore, we proceed to the active learning stage to unmask the broader variation in recruitment tactics across geographies.

3.2. Active Learning

The process designed for generating a training set provides us with initial well-balanced training data. However, it is clearly biased by the purview of domain experts and does not provide a complete view of the numerous styles/types of recruiting posts on the deep web. Thus, we use pool-based active learning, which is known to improve classifiers with significantly reduced manual labeling

effort (Settles 2012, 2009). Rather than labeling a random subset of posts, these approaches direct costly labeling effort towards posts that are estimated to resolve the most uncertainty (i.e., improve the accuracy) of the current classifier. In particular, we begin by training an initial deep neural network (which has shown great success in text classification tasks (Otter et al. 2020)) using the initial training set of 1651 posts. We then use this classifier to assign a prediction probability to each unlabeled post on how likely it is to be recruiting-related – this metric captures the prediction uncertainty that is traditionally used by active learning to prioritize labeling (Zhu et al. 2010).

However, as noted earlier, our active learning objective is not simply to maximize the accuracy of our classifier across all posts (which would have the consequence of focusing labeling efforts on locations with many posts), but to uncover an accurate representation of the underlying network across locations. We address this objective mismatch by incorporating geographical diversity and the likelihood of identifying new network connections in our learning procedure. Specifically, we add two additional metrics to our active learning objective: (i) a 'node information' score that prioritizes posts in under-sampled locations that may have additional recruitment activity, and (ii) an 'edge information' score that prioritizes posts corresponding to an under-sampled pair of locations (as determined by the meta data) that may represent a new inferred trafficking route. Our algorithm uses this objective to prioritize a batch of unlabeled posts for labeling. The resulting batch of labeled posts are then added to the labeled training data, and the deep learning network is re-trained. This active learning process is repeated until the model performance converges. Overall, we obtained labels on approximately 50,000 posts, identifying approximately 7000 recruiting-related posts. Despite the heavy data imbalance, this corresponds to 14% of the labels being positive. Furthermore, our active learning process allowed us to uncover 27 different types of recruiting tactics, far outperforming the initial expert-identified vocabulary which only identified 3 types of recruitment tactics.

We first define some notation. Let X be the pool of all posts; at any point of time, this pool is composed of mutually exclusive sets $X = X_0 \cup X_1$ where X_0 is the set of unlabeled posts and X_1 is the (much smaller) set of labeled posts with corresponding binary labels Y_1. Each post $x \in X$ is associated with two quantities: a (potentially empty) set of locations L_x and a (potentially empty) set of identifying information M_x (e.g., phone number, email). 94% of posts in our sample have at least one location and 69% of posts have at least one identifier.

Then, for every unlabeled post $x \in X_0$, we can construct the set of potential "edges" (i.e., between a pair of locations) that it may inform for network discovery. We are specifically interested in edges in the commercial sex supply chain network that may carry trafficking risk. Thus, we define the set:

$$E_x = \{\ell_1 \Leftrightarrow \ell_2 \mid \exists x' \in X \quad s.t. \quad M_x \cap M_{x'} \neq \varnothing \quad and \quad \ell_1 \in L_x\,, \ell_2 \in L_{x'}\}.$$

In other words, for any unlabeled post $x \in X_0$, E_x captures the number of potential recruiting-sales or recruiting-recruiting location edges we will identify (based on some shared identifier from the meta data M_x) if x is found to be a recruiting post. Note that some of these edges may already be known to carry (or likely not carry) trafficking risk based on other labeled samples.

For each batch in active learning, we iteratively re-train our selected model using the currently labeled posts (X_1, Y_1) as our training set; this yields a model $f : x \to (0, 1)$ that predicts the likelihood that a post x is a recruiting post based on its text. Then, we apply the model to predict the probability $f(x)$ that each currently unlabeled post $x \in X_0$ is a recruiting post. Traditional active learning would solely rely on this metric to determine which posts to prioritize for labeling – specifically, we define the function:

$$\chi(x) \ = 1 - |\, \frac{1}{2} - f(x) \,|\, .$$

$\chi(\cdot)$ captures the uncertainty of a post's prediction. Traditional active learning seeks to reduce labeling effort by focusing effort away from posts that already have confident predictions (e.g., clearly sales, $f(x) = 0$, or clearly recruitment, $f(x) = 1$) and are therefore unlikely to improve the accuracy of the current predictive model. Instead, active learning prioritizes posts $x \in X_0$ that have high values of $\chi(x)$ (i.e., values of $f(x)$ that are close to 1); these are the posts for which the current predictive model is relatively uninformative, and therefore augmenting the training set with the labels of these posts may improve the accuracy of the model.

However, such an approach focuses purely on improving the predictive accuracy of f across all posts. As noted earlier, our objective is more nuanced – we seek to uncover an accurate representation of the underlying supply chain network across locations. We address this objective mismatch by incorporating geographical diversity and the likelihood of identifying new network edges in our learning procedure. Unlike traditional active learning, our new prioritization will crucially rely on the meta data (L_x, M_x) associated with a post x.

Thus, in addition to $\chi(\cdot)$, we define additional metrics – the 'node uncertainty' and the 'edge uncertainty' to capture how a post contributes to geographically diverse coverage. To formalize these metrics, we require some additional notation. We begin by defining two useful subsets of unlabeled posts:

$$\Delta = \{x \in X_0 \mid 0.4 \leq f(x) \leq 0.8\},$$

$$\eta = \{x \in X_0 \mid f(x) > 0.8\}.$$

Δ captures uncertain posts, while η captures likely-recruitment posts. The upper and lower bounds are tuning parameters that were chosen to optimize the performance of the active learning procedure.

Node Uncertainty. Let the set of unlabeled posts corresponding to a certain location be denoted by

$$V(\ell) = \{x \in X_0 \mid \ell \in L_x\}.$$

Then for a given location ℓ, we define the 'node uncertainty' to be:

$$N(\ell) = \frac{|\,\Delta \cap V(\ell)\,|^2}{|\,\eta \cap V(\ell)\,| + 1}.$$

$N(\ell)$ captures the extent to which we distinguish potential recruitment tactics at location ℓ. Specifically, if the numerator (number of uncertain posts associated with location ℓ) is high, we wish to prioritize posts associated with this location; in contrast, if the denominator is high (we have identified many likely-recruitment posts already), we wish to de-prioritize associated posts. Then, for every unlabeled post $x \in X_0$, we can compute a normalized score of how much labeling it may contribute to reducing node uncertainty for its set of locations L_x:

$$N(x) = \frac{1}{|\,L_x\,|} \sum_{\ell \in L_x} N(\ell).$$

Edge Uncertainty. Analogously, let the set of unlabeled posts corresponding to a certain pair of locations be denoted by

$$T(e) = \{x \in X_0 \mid e \in E_x\}.$$

Then, for a given edge e between a pair of locations, we define the 'edge uncertainty' to be

$$M(e) = \frac{|\,\Delta \cap T(e)\,|^2}{|\,\eta \cap T(e)\,| + 1}.$$

$M(e)$ captures the extent to which we distinguish potential recruitment-to-sales or recruitment-recruitment pathways in the supply chain network for an edge e between a pair of locations. If the numerator (number of uncertain posts associated with edge e) is high, we wish to prioritize posts associated with this location; in contrast, if the denominator is high (we have identified many likely-recruitment posts already), we wish to de-prioritize associated posts. Then, for every unlabeled post $x \in X_0$, we can compute a normalized score of how much labeling it may contribute to reducing edge uncertainty for its set of locations E_x:

$$M(x) = \frac{1}{|\,E_x\,|} \sum_{e \in E_x} M(e).$$

Active Learning Strategy. Our active learning strategy proceeds in batches. In each batch, we use the current predictive model f to make predictions on every currently unlabeled post $x \in X_0$ (several predictive models were compared prior to selecting the deep neural net architecture

used, please see Appendix A for details). All likely-recruitment posts are automatically prioritized for labeling. Following traditional active learning, we also prioritize posts with high prediction uncertainty $\chi(x)$. Then, to improve network discovery, we prioritize posts that have a high score $N(x)$ for reducing node uncertainty, and a high score $M(x)$ for reducing edge certainty in our supply chain network. We rank the unlabeled posts in $x \in X_0 \cap \eta^c$ according to our modified metrics and choose the top 4,000 posts to label. Once these labels are obtained, we appropriately modify X_0, X_1 and retrain our deep learning model f on the augmented training set X_1. We then re-compute our set of unlabeled likely-recruitment posts ; we stop the active learning process when this set is empty (see Algorithm 1).

Algorithm 1 Active Learning Pseudocode

Input: unlabeled posts X_0, labeled posts X_1, initial model f trained on initial training set
Predict $f(x)$ for every $x \in X_0$
Compute the set of unlabeled likely recruitment posts η
while $\eta \neq \varnothing$ **do**
    Initialize prioritized posts for labeling to $B = \eta$
    Compute 'prediction uncertainty' $\chi(x)$ for every remaining $x \in X_0 \cap \eta^c$
    Compute 'node uncertainty' $N(\ell)$ for every location ℓ
    Compute 'edge uncertainty' $E(e)$ for every edge e
    Compute $N(x), M(x)$ for every remaining $x \in X_0 \cap \eta^c$
    Sort remaining posts $x \in X_0 \cap \eta^c$ by descending order of $\chi(x)$ and then by descending order of $N(x), M(x)$
    Select top 4000 posts P and add to batch to be labeled $B \leftarrow B \cup P$
    Obtain manual labels (x, y) for all $x \in B$
    Update labeled set $X_1 \leftarrow X_1 \cup B$, and unlabeled set $X_0 \leftarrow X_0 \cap B^c$
    Train new predictive model $f(\cdot)$ using augmented training data (X_1, Y_1)
    Compute set of unlabeled likely recruitment posts η
end

We ran 13 batches of active learning. Figure 4 shows the histogram of prediction scores $f(x)$ on our unlabeled posts X_0 after each batch. In the first batch (after training on only our initial training set), we observe a very large spread of prediction scores across the interval (0,1), indicating a large degree of uncertainty. In later batches, as we iteratively label both likely-recruitment and uncertain posts, we observe that the number of likely-recruitment posts (i.e., predictions above 0.8) among the unlabeled set X_0 decreases steeply as the results converge (note the scale of the y-axis

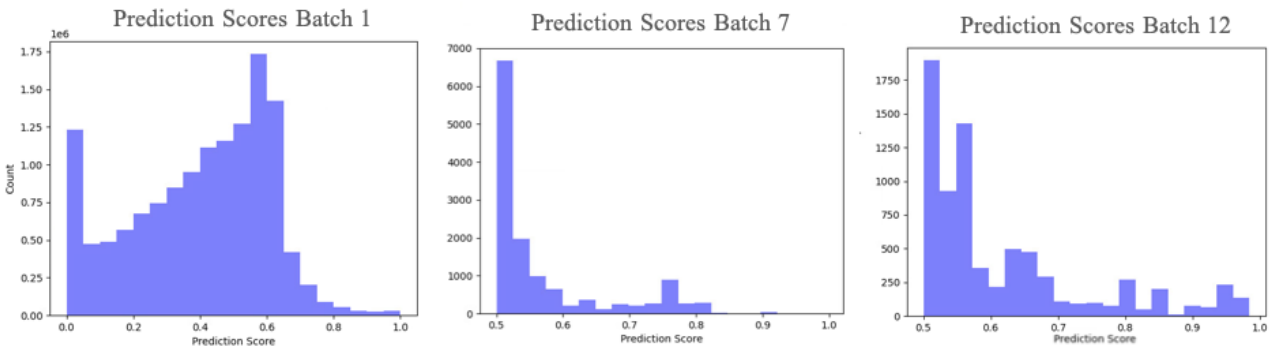


Figure 4 **Prediction scores from our model on the unlabeled data for three batches across the process. Note that the y-axis and x-axis are different across plots. Batch 12 has far fewer uncertain posts (prediction score above .5) than any prior batch.**

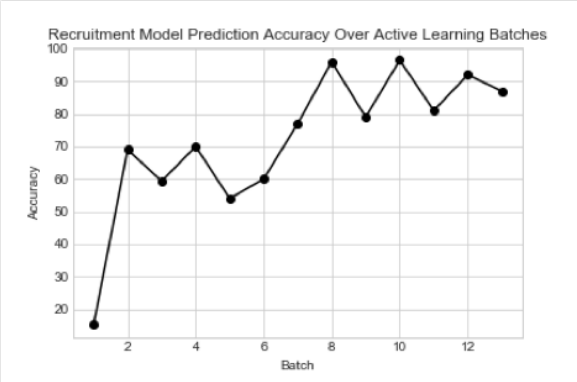


Figure 5 **Accuracy of the predictive model f across 13 batches of the active learning process.**

across batches in Figure 4). Figure 5 shows how the accuracy of f evolves with each additional batch; note that its performance asymptotes, suggesting convergence.

Throughout the process, we obtained labels on a total of 50,199 total posts, of which 6,953 posts were identified as recruitment. This corresponds to 14% of the labels being positive, as opposed to <0.1% if we had labeled randomly. Critically, while our initial training set only identified 3 types of recruitment tactics, our active learning process uncovered 27 recruitment tactics (see Appendix B), demonstrating the effectiveness of our proposed approach. We additionally note that relying on traditional active learning alone would have directed our labeling efforts to posts from large cities (where the majority of posts occur), missing out on key recruitment tactics we identify in smaller cities (where we actually find recruitment dominates).

Network Creation. Finally, we connect the identified recruitment and sales posts using shared meta data to determine which posts were made by the same entity (see last panel of Figure 2). Along with the locations of the posts, this allows us to identify the geographic network connections

underlying commercial sex supply chains. We use the following variables extracted from the meta data of posts to identify entities: email, phone number, username, URL, and social media handle. We find 43,521 connections in total from recruitment to commercial sex sales posts; surprisingly, 10% of recruitment posts account for 85% of the connections.

4. Results

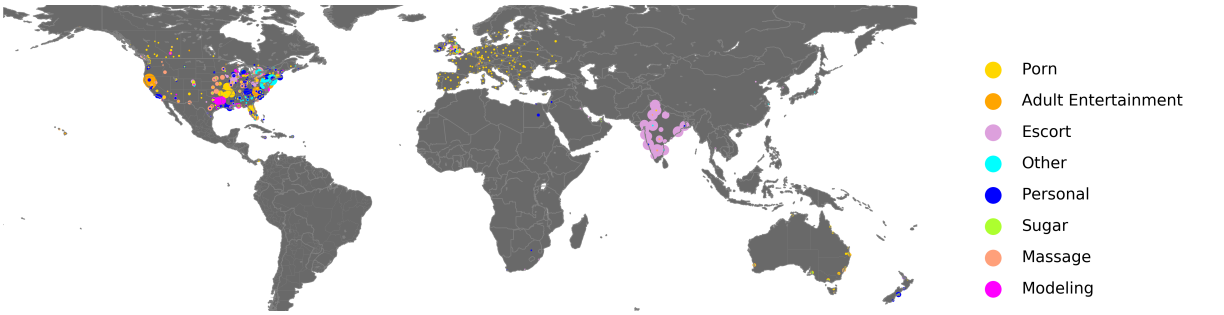
Recruitment Activity. Figure 6 shows the global map of recruitment hotspots and the types of recruitment tactics identified in our data. Note that recruitment posts in the 'escort' category indicate potential sex work, while posts in all other categories do not indicate sex work. As shown in Figure 6a, we found significant activity primarily in the United States, Canada, Europe, India and Australia; this is likely due to our restriction to posts in the English language (see discussion in §5).

We observe significant geographic variation in the approaches used to recruit victims (see Figures 6a and 6b); the full list of recruitment types is in Appendix B. For example, within the United States, individuals are primarily targeted through modeling and porn offers in the Midwest, escort and adult entertainment services (e.g., strip clubs) in the East and West coasts, and even personal ads in several major cities. More globally, victims are targeted primarily through porn and adult entertainment offers in Europe, and escort services in India. Early interventions for preventing exploitation of vulnerable populations have recommended 'job search' training to educate potential victims on the risks associated with responding to different types of recruitment posts (Murphy 2016). These results can be used to tailor such educational programs towards the currently popular recruitment approaches in those specific locales.

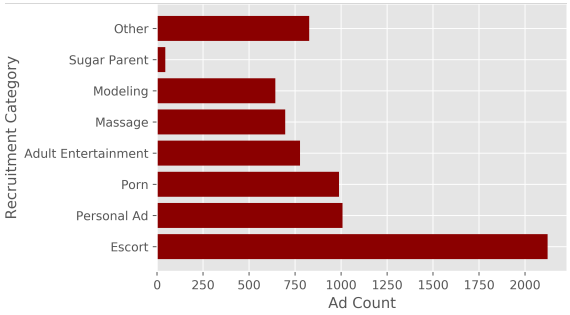
We also construct recruitment-recruitment networks, creating an undirected edge between a pair of locations if they each have a high volume of recruiting posts (at least 150) by the same entity. Figure 6c shows the resulting network within the United States. We observe that many recruiters operate in multiple locations spanning large distances, suggesting a highly organized effort.

Inferring Human Trafficking Risk. As discussed earlier, linking deceptive (non-sex) recruitment offers to commercial sex sales by the same entity strongly suggests that trafficking may have occurred. Figure 7a shows an example connection between an identified recruitment post and two sales posts with shared meta data; although the recruiting post offers a modeling employment opportunity in Canada, the same phone number appears in over a hundred sex sales posts in Canada, the United Kingdom, the United States and Australia. This suggests that the modeling post is a masked attempt to recruit victims into an international sex trafficking organization.

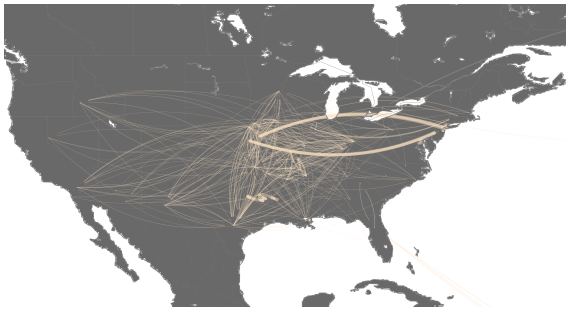
To study trafficking risk at scale, we construct recruitment-to-sales pathways: we create a directed edge between a pair of locations if there is a recruiting post and a sales post with shared meta



(a) Recruitment Activity by Category Uncovered by Active Learning Algorithm



(b) Types of Recruitment



(c) Connections amongst Recruiters

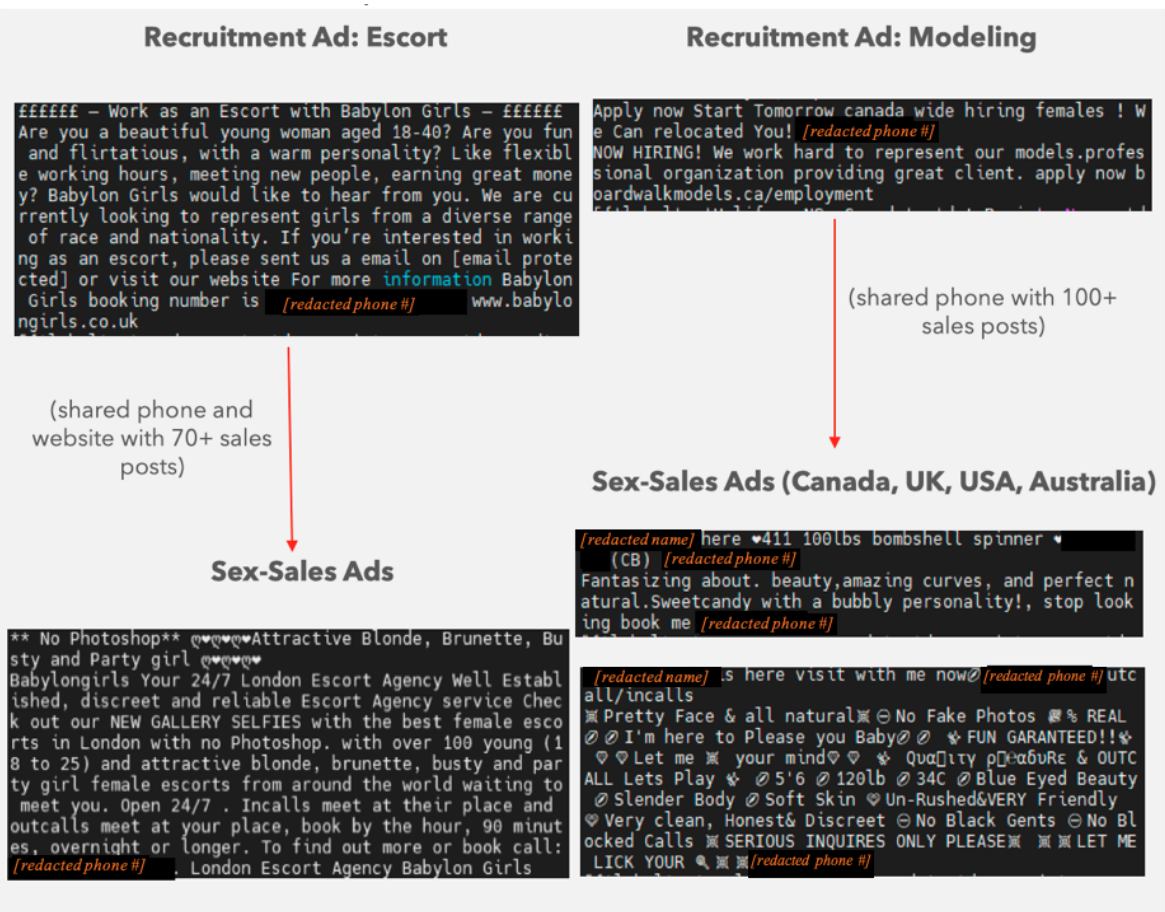
Figure 6 **The top panel (a) shows recruitment hotspots and tactics identified in our data. Larger markers indicate more posts. The bottom left panel (b) shows the histogram of recruitment posts by category across the world. The bottom right panel (c) shows the recruitment-recruitment network in the United States. We display an edge between a pair of locations if there are at least 150 recruitment posts that share meta-data (thus, are posted by the same entity); thicker lines indicate more recruitment posts (capped at 2000 posts for visual clarity).**

data. Figure 7b shows the resulting commercial sex supply chain, restricting to edges that have at least 150 occurrences. Importantly, we find that over 95% of these connections are accounted for by deceptive recruitment posts that do not mention any potential for sex work.

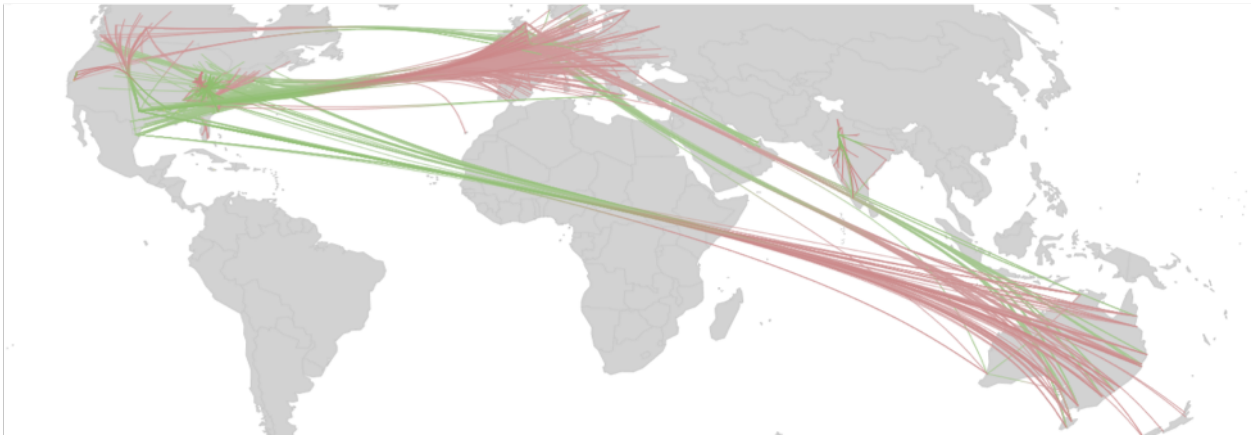
We also find that 10% of recruitment ads are responsible for 85% of edges in the supply chain network, suggesting that there are a few large-scale entities driving a significant portion of trafficking activity. This result underscores the importance of our modified active learning procedure, which targets network discovery in addition to the traditional objective of improving classification accuracy.

Figure 8 delves further into the domestic recruitment-to-sales supply chain connections identified by our analysis. Domestic network connections are most prominent in the United States (Figure 8a) and India (Figure 8b).

Recruitment vs. Sales Pressure. We distinguish 'sender' cities (where victims are recruited) from 'receiver' cities (where sex sales occur). For example, in India, we observe that recruitment occurs in coastal locations, while sales primarily occur in the capital, New Delhi (see Fig. 4B).

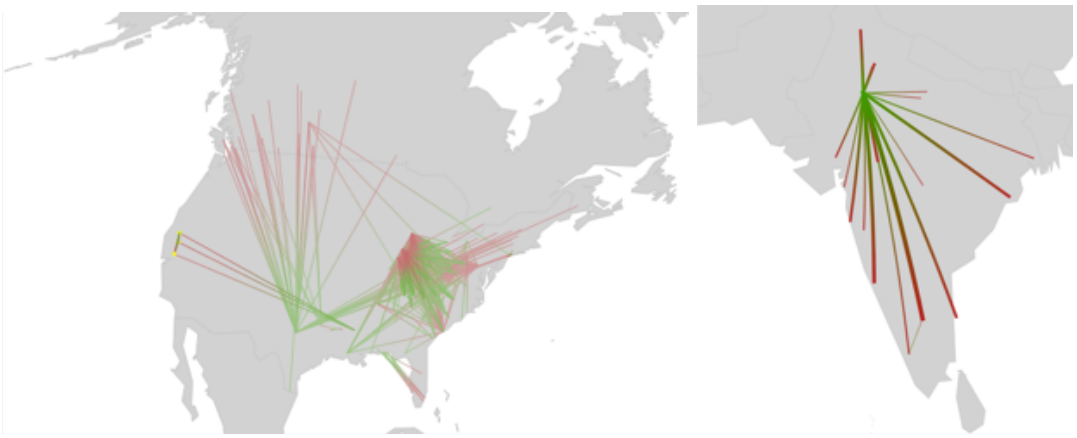


(a) Example Recruitment-Sales Connections



(b) Recruitment to Sex-Sales Pathways Unmasked

Figure 7 **The top panel (a) shows example escort and modeling recruitment posts uncovered in the UK and Canada that share the same phone number as sex sales posts in Canada and other countries; note that we have redacted personal identifiable information with square brackets and the type of information (e.g., [redacted phone #]). The bottom panel (b) shows the resulting global view of trafficking risk in commercial sex supply chain networks from deceptive recruitment offers (red) to commercial sex sales (green) unmasked from our algorithm. Network is restricted to edges with at least 100 occurrences.**



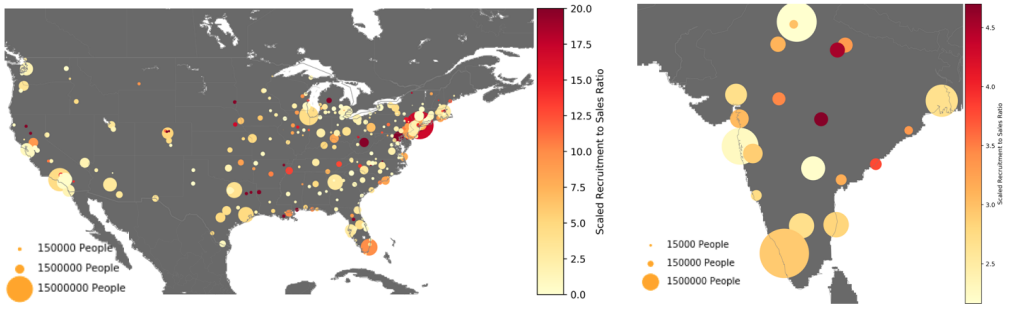
(a) USA Recruitment- Sex Sales Network

(b) India Recruitment- Sex Sales Network

Figure 8 **Closer examination of inferred likely trafficking routes in the United States (a) and India (b). Network shows pathways from recruitment offers (red) to commercial sex sales (green) with at least 100 occurrences.**

Similarly, in the United States, recruitment is concentrated in suburban locations (e.g., Scranton, Redding), while sales primarily occur in major cities (e.g., Miami, New York City, Los Angeles). Figure 9a shows a map of relative recruitment to sales pressure across the United States; we observe that densely populated locations tend to be receiver cities while less populated locations tend to be sender cities. Note that relying on traditional active learning would have directed our labeling efforts to posts from large cities (where the majority of posts occur), missing out on key recruitment hotspots in smaller cities identified by our modified active learning procedure. These results can be used to tailor interventions in specific locales, e.g., invest in education and social work to reduce recruitment in sender cities, and invest in law enforcement to prosecute sex sales in receiver cities.

We also examine the characteristics of top 50 sender and receiver cities identified in the United States; only 17 of these locations overlap, underscoring that recruitment and sex sales are typically concentrated in distinct locations. Using census data, we find that sender cities tend to be more economically constrained (have higher poverty rates and lower household incomes), and furthermore have higher crime rates relative to receiver cities; details of this analysis are provided in Appendix C. These results suggest that sender cities may not have as many resources as larger receiver cities to prevent trafficking of their vulnerable populations; thus, they may benefit from collaborations with (better-funded) counter-trafficking agencies in larger receiver cities. Such collaborations may be particularly valuable when there is a likely recruitment-to-sales trafficking route between the two cities. For example, we identified an entity that frequently recruits (deceptively) in Redding, CA and sells sex in Sacramento, CA; therefore, a collaboration between agencies in Redding and Sacramento



(a) USA Recruitment- Sex Sales Pressure Ratios (b) India Recruitment- Sex Sales Pressure Ratios

Figure 9 **Map of relative recruitment to sales pressure across locations in the United States (A) and India (B). Color represents the ratio of recruitment over sales ads from the deep web scaled by a factor of 10,000 due to the substantial difference in activity levels. The size of the bubbles corresponds to population size of the city, highlighting that smaller cities tend to have higher recruitment pressure (in dark red) and larger cities have higher sex sales activity (light yellow).**

would simultaneously provide support for the smaller and more economically constrained Redding population, and enable targeting of a potential trafficking entity from both ends of its supply chain.

Other Datasets. To the best of our knowledge, our study is the first to characterize recruitment at scale in commercial sex, which allows us to uniquely infer trafficking risk. However, we can compare our results on sex sales from the deep web against other sources. Specifically, we consider Rubmaps.ch (a popular review site for massage parlors with sexual services) as well as suspicious businesses identified through Google Places. Details are provided in Appendix C. We find that commercial sex sales activity identified on the deep web roughly aligns with activity identified through other sources, but recruitment activity is distinct and uniquely identified by our analysis. Thus, we provide the first large-scale network view of trafficking risk in commercial sex supply chains, from recruitment to sales.

5. Discussion

We leverage machine learning and deep web data to construct the first large-scale and data-driven view of commercial sex supply chains. Our approach uniquely allows us to link deceptive recruitment activity to sex sales by the same entity to unmask trafficking risk. These results yield several policy-relevant insights.

First, inferring likely recruitment-to-sales pathways can help law enforcement agencies along potential trafficking routes better coordinate efforts. The FBI reports that the most effective way to investigate human trafficking is through a "collaborative, multi-agency approach with our federal, state, local, and tribal partners" (FBI 2021). For example, they hold an annual week-long counter-trafficking 'sweep,' where law enforcement officials across the United States respond undercover to

sex sales posts to generate leads on traffickers. This synchronized effort has shown great success, leading to 67 arrests in 2019 (FBI 2019), but it has its drawbacks. Naturally, a sustained counter-trafficking effort would be more effective; however, it is costly for many agencies to simultaneously collaborate in this fashion, and there is currently no systematic way to determine which collaborations to prioritize (Shively et al. 2012). Also, sweeps are focused on major cities with high sex sales pressure, largely ignoring high-risk suburban locations with high recruitment pressure. Our analysis uncovers likely trafficking routes that can help prioritize partnerships between impacted law enforcement jurisdictions; moreover, instead of focusing purely on sex sales, these collaborations can holistically tackle an entity's trafficking supply chain, from recruitment to sales.

Second, identifying region-specific exploitative behaviors can inform targeted local policies and interventions. Social policy plays an important role in preventing vulnerable victims from being trafficked (Orme and Ross-Sheriff 2015), as well as rehabilitating victims after their trafficking experience (Rafferty 2008). While the latter (mitigation) is more prevalent, the former (prevention) shows significant promise since many victims are domestic, e.g., an estimated 67% of trafficking victims in the United States are United States citizens (Jorgensen and Sandoval 2019), and 93% of victims in Canada are Canadian citizens (Lopez-Martinez 2020). To this end, our results provide large-scale insight into where and how victims are (often deceptively) recruited. Cities with high recruitment pressure may prefer to focus their resources on preventative measures and can furthermore tailor interventions towards the recruitment tactics frequently seen in their specific locale. Prioritizing resource allocation to maximize impact in this manner is valuable since social resources are often highly constrained.

There are some limitations that may materialize if there is significant adoption of these methods in counter-trafficking. First, criminals may respond by creating new recruitment templates in order to evade detection. This can be combatted by periodically re-training the machine learning model using our active learning approach and ensuring up-to-date coverage of commercial sex websites. Second, sex trafficking entities may cease using the same contact information (i.e., meta data) across locations, making it more difficult to infer an organization's recruitment-to-sales pathways (although one can still reliably infer recruitment and sales pressure). In this case, new methods can be explored for mappings, e.g., based on shared post verbiage/style; these methods have already shown success identifying sex sales ads by the same organization (Li et al. 2018, Dubrawski et al. 2015). We note that it is unlikely that criminals will respond with these shifts in the near term. Finally, while deep web data provides a significant opportunity to scale the collection of information, it may fail to provide adequate coverage of some vulnerable populations. For instance, half of the cases reviewed by the UNODC in 2020 used the internet (Kangaspunta et al. 2020), but there is still a significant amount of trafficking activity conducted offline (e.g., through word-of-mouth).

Relatedly, our choice of websites (informed by law enforcement partners) as well as limitation to the English language may limit visibility of illicit activity occurring elsewhere or in local languages. As discussed in Section 2, a promising direction of future work is adapting our approach to other languages to improve global coverage. Thus, it is important to note that any insights from our approach should complement rather than replace traditional leads (e.g., survivor interviews, prior case outcomes, etc.), which may provide better coverage over vulnerable populations that are underrepresented by our analysis.

This work demonstrates how powerful machine learning tools can be applied in tandem with domain expertise for inference in settings with highly imbalanced and networked data. Our approach can be leveraged to investigate other type of trafficking with a heavy web presence (e.g., drugs, weapons, etc.) or, more broadly, in applications that require uncovering granular local patterns from large-scale, unstructured textual data.

Acknowledgments

The authors are grateful to Chris Dickson and Danielle Smalls of Uncharted Software for assistance with curating and contextualizing the core deep web dataset, as well as Carolina Holderness and Pierre Griffith of the Human Trafficking Response Unit at the Manhattan District Attorney's Office for providing invaluable domain insights. The authors also thank Tsai-Hsuan Chung for collecting auxiliary data in support of this work, and David Jonker for helpful comments on an earlier draft.

References

Albright E, D'Adamo K (2017) Decreasing human trafficking through sex work decriminalization. *AMA journal of ethics* 19(1):122–126.

Androff DK (2011) The problem of contemporary slavery: An international human rights challenge for social work. *International Social Work* 54(2):209–222.

BEA (2018) U.s. bureau of economic analysis. URL `https://apps.bea.gov/regional/downloadzip.cfm`.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57(1):289–300.

BLS (2018) U.s. bureau of labor statistics. URL `https://www.bls.gov/lau/#tables`.

Bouche V, Crotty SM (2018) Estimating demand for illicit massage businesses in houston, texas. *Journal of human trafficking* 4(4):279–297.

Census (2018) U.s. census. URL `https://data.census.gov/cedsci/`.

Chan H, Tran-Thanh L, Wilder B, Rice E, Vayanos P, Tambe M (2018) Utilizing housing resources for homeless youth through the lens of multiple multi-dimensional knapsacks. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 41–47.

Chen XW, Lin X (2014) Big data deep learning: challenges and perspectives. *IEEE access* 2:514–525.

Dank ML, Khan B, Downey PM, Kotonias C, Mayer D, Owens C, Pacifici L, Yu L (2014) Estimating the size and structure of the underground commercial sex economy in eight major us cities .

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Diaz M, Panangadan A (2020) Natural language-based integration of online review datasets for identification of sex trafficking businesses. *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, 259–264 (IEEE).

Dligach D, Palmer M (2011) Good seed makes a good crop: accelerating active learning using language modeling. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 6–10.

Do CB, Ng AY (2005) Transfer learning for text classification. *Advances in Neural Information Processing Systems* 18:299–306.

Dubrawski A, Miller K, Barnes M, Boecking B, Kennedy E (2015) Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking* 1(1):65–85.

FBI (2016) U.s. department of justice. URL `https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-8/table-8.xls/view`.

FBI (2019) Operation independence day,. URL `https://www.fbi.gov/news/stories/operation-independence-day-2019`.

FBI (2021) Human trafficking,. URL `https://www.fbi.gov/investigate/violent-crime/human-trafficking`.

Flynn C, Alston M, Mason R (2014) Trafficking in women for sexual exploitation: Building australian knowledge. *International Social Work* 57(1):27–38.

Gensim (2021) Gensim 4.0.1. URL `https://www.snorkel.org/`.

Gonsior J, Thiele M, Lehner W (2020) Weakal: Combining active learning and weak supervision. *International Conference on Discovery Science*, 34–49 (Springer).

Hall E, Dickson C, Schroh D, Wright W (2015) Tellfinder: Discovering related content in big data (VIS).

Heilemann T, Santhiveeran J (2011) How do female adolescents cope and survive the hardships of prostitution? a content analysis of existing literature. *Journal of Ethnic & Cultural Diversity in Social Work* 20(1):57–76.

Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural computation* 9(8):1735–1780.

Hodge DR, Lietz CA (2007) The international sexual trafficking of women and children: A review of the literature. *Affilia* 22(2):163–174.

HUD (2019) Us department of housing and urban development. URL `https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/`.

ILO (2017) Human trafficking by the numbers.

Johnson BC (2012) Aftercare for survivors of human trafficking. *Social Work & Christianity* 39(4).

Jones L, Engstrom DW, Hilliard T, Diaz M (2007) Globalization and human trafficking. *J. Soc. & Soc. Welfare* 34:107.

Jorgensen S, Sandoval P (2019) Experts: Trump's tape bound women trafficking claim is misleading,. *CNN* URL `https://www.cnn.com/2019/01/27/us/human-trafficking-fact-check/index.html`.

Kangaspunta K, Sarrica F, Serio G, Kelly W, Samson J, Wills C (2020) Global report on trafficking in persons 2020 .

Kaya YB, Maass KL, Dimas GL, Konrad R, Trapp AC, Dank M (2022) Improving access to housing and supportive services for runaway and homeless youth: Reducing vulnerability to human trafficking in new york city. *arXiv preprint arXiv:2202.00138* .

Kejriwal M, Kapoor R (2019) Network-theoretic information extraction quality assessment in the human trafficking domain. *Applied Network Science* 4(1):1–26.

Keskin BB, Bott GJ, Freeman NK (2021) Cracking sex trafficking: Data analysis, pattern recognition, and path prediction. *Production and Operations Management* 30(4):1110–1135.

Konrad R, Maass KL, Trapp AC (2020) A perspective on how to conduct responsible anti-human trafficking research in operations and analytics. *arXiv preprint arXiv:2006.16445* .

Konrad RA, Trapp AC, Palmbach TM, Blom JS (2017) Overcoming human trafficking via operations research and analytics: Opportunities for methods, models, and applications. *European Journal of Operational Research* 259(2):733–745.

Kosmas D, Sharkey TC, Mitchell JE, Maass KL, Martin L (2020) Interdicting restructuring networks with applications in illicit trafficking. *arXiv preprint arXiv:2011.07093* .

Kotrla K (2010) Domestic minor sex trafficking in the united states. *Social work* 55(2):181–187.

Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 5(4):221–232.

Laczko F (2002) Human trafficking: the need for better data. *Migration Information Source* 1:61–80.

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

Li L, Simek O, Lai A, Daggett M, Dagli CK, Jones C (2018) Detection and characterization of human trafficking networks using unsupervised scalable text template matching. *2018 IEEE International Conference on Big Data (Big Data)*, 3111–3120 (IEEE).

Li R, Tobey M, Mayorga M, Caltagirone S, zaltın O (2021) Detecting human trafficking: Automated classification of online customer reviews of massage businesses. *Available at SSRN 3982796* .

Lin M, Chen Q, Yan S (2013) Network in network. *arXiv preprint arXiv:1312.4400* .

Liu G, Guo J (2019) Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing* 337:325–338.

Lopez-Martinez M (2020) Sex trafficking still a prevalent issue across canada, advocates and police say. *CTVNews* URL `https://www.cnn.com/2019/01/27/us/human-trafficking-fact-check/index.html`.

Maass KL, Trapp AC, Konrad R (2020) Optimizing placement of residential shelters for human trafficking survivors. *Socio-Economic Planning Sciences* 70:100730.

McCallumzy AK, Nigamy K (1998) Employing em and pool-based active learning for text classification. *Proc. International Conference on Machine Learning (ICML)*, 359–367 (Citeseer).

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J (2021) Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)* 54(3):1–40.

Murphy LT (2016) Labor and sex trafficking among homeless youth. *A Ten City Study (Executive Summary)* .

Nashaat M, Miller J (2021) Improving news popularity estimation via weak supervision and meta-active learning. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 2679.

NCES (2017) National center for education statistics. URL `https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?goToReportId=7`.

Nwankpa C, Ijomah W, Gachagan A, Marshall S (2018) Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378* .

Office GA (2006) Human trafficking: Better data, strategy, and reporting needed to enhance us antitrafficking efforts abroad. *Trends in Organized Crime* 10:16–38.

Okech D, Choi YJ, Elkins J, Burns AC (2018) Seventeen years of human trafficking research in social work: A review of the literature. *Journal of evidence-informed social work* 15(2):103–122.

Olsson F (2009) A literature survey of active machine learning in the context of natural language processing .

Orme J, Ross-Sheriff F (2015) Sex trafficking: Policies, programs, and services. *Social work* 60(4):287–294.

Otter DW, Medina JR, Kalita JK (2020) A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* .

Potocky M (2010) The travesty of human trafficking: A decade of failed us policy. *Social Work* 55(4):373–375.

Proximity (2009) Proximity. URL `http://proximityone.com/metro_healthinsurance.htm`.

Raets S, Janssens J (2019) Trafficking and technology: Exploring the role of digital communication technologies in the belgian human trafficking business. *European Journal on Criminal Policy and Research* 1–24.

Rafferty Y (2008) The impact of trafficking on children: Psychological and social policy perspectives. *Child Development Perspectives* 2(1):13–18.

Rajapakse T (2020) Simpletransformers. URL `https://simpletransformers.ai/`.

Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C (2017a) Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, 269 (NIH Public Access).

Ratner AJ, Ehrenberg HR, Hussain Z, Dunnmon J, Ré C (2017b) Learning to compose domain-specific transformations for data augmentation. *Advances in neural information processing systems* 30:3239.

Rehurek R, Sojka P (2010) Software framework for topic modelling with large corpora. *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (Citeseer).

Roby JL (2005) Women and children in the global sex trade: Toward more effective policy. *International Social Work* 48(2):136–147.

Roby JL, Vincent M (2017) Federal and state responses to domestic minor sex trafficking: the evolution of policy. *Social work* 62(3):201–210.

Rong X (2014) word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738* .

Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45(11):2673–2681.

Settles B (2009) Active learning literature survey .

Settles B (2012) Active learning (synthesis lectures on artificial intelligence and machine learning)(morgan and claypool publishers, san rafael, ca) .

Shelters W (2021) Women's shelters. URL `https://www.womenshelters.org/#state_list`.

Shively M, Kliorys K, Wheeler K, Hunt D (2012) A national overview of prostitution and sex trafficking demand reduction efforts .

Smirnov NV (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou* 2(2):3–14.

Snorkel (2021) Snorkel. URL `https://www.snorkel.org/`.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1):1929–1958.

Symeonidis S, Effrosynidis D, Arampatzis A (2018) A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications* 110:298–310.

TellFinder (2021) Tellfinder alliance: a global counter-human trafficking partner network, empowered by data,. URL `https://www.tellfinderalliance.com/`.

TensorFlow (2021) Tensorflow. URL `https://www.tensorflow.org/`.

UNODC (2020) Unodc human trafficking case law database .

Walker-Rodriguez A, Hill R, Trafficking HS (2011) Fbi law enforcement bulletin.

Witte M (2018) The anti-trafficking movement needs better data to solve the problem, stanford researchers say. *Stanford News Service* URL `https://news.stanford.edu/press-releases/2018/09/05/get-good-data-human-trafficking/`.

Yang Y, Ma Z, Nie F, Chang X, Hauptmann AG (2015) Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision* 113(2):113–127.

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV (2019) Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237* .

Zhang C, Ré C, Cafarella M, De Sa C, Ratner A, Shin J, Wang F, Wu S (2017) Deepdive: Declarative knowledge base construction. *Communications of the ACM* 60(5):93–102.

Zhou ZH (2018) A brief introduction to weakly supervised learning. *National science review* 5(1):44–53.

Zhu J, Li L, Jones C (2019) Identification and detection of human trafficking using language models. *2019 European Intelligence and Security Informatics Conference (EISIC)*, 24–31 (IEEE).

Zhu J, Wang H, Hovy E, Ma M (2010) Confidence-based stopping criteria for active learning for data annotation. *ACM Transactions on Speech and Language Processing (TSLP)* 6(3):1–24.

Appendix A: Predictive Model Selection

Before proceeding to our active learning strategy, we must select a machine learning model for prediction. Deep neural networks (DNNs) have shown great success in text classification tasks (Minaee et al. 2021), but there are a number of state-of-the-art approaches that may be promising. Thus, we train and evaluate 6 types of DNN models using our initial training data: 4 rely only on our data alone, while 2 additionally incorporate transfer learning from existing language models.

To improve the quality of our initial predictive model, we also augment our initial training data by adding structured noise to the labeled posts. We leverage a series of transformation functions that replace names, adjectives, and verbs with synonyms in order to generate a set of synthetic labeled posts; such an approach is helpful when the training set is small because it helps the predictive model avoid overfitting to irrelevant features (e.g., names) (Ratner et al. 2017b). However, as we collect additional data through active learning, we discard the synthetic posts generated by data augmentation in model training.

We reserve a 20% random subsample of our initial training data as a validation set, on which we evaluate the predictive quality of all 6 models (see results in Table 2). The first four models are built using Keras in Tensorflow (TensorFlow 2021). The base model ("Model 1") takes an input of tokenized sequences that represent each post. First, the input enters an embeddings layer that allows the model to modify the word-vectors used to encode the text during model training while it learns which posts are likely recruitment. The learned embeddings are then fed into a global average pooling layer to help prevent overfitting (Lin et al. 2013). The final layer is a densely connected layer with a sigmoid activation function, which is useful for predicting probabilities (Nwankpa et al. 2018). Our second model ("Model 2") additionally includes dropout (a regularization method in which some number of nodes in the deep neural network are ignored during training), which has been shown to reduce overfitting to the training set (Srivastava et al. 2014). We also include a bias initializer to help address the remaining data imbalance in our training set (Krawczyk 2016). Next, we test two long short-term memory (LSTM) models (a type of recurrent neural network that is capable of learning the order dependence in a sequence), which are useful for text classification (Liu and Guo 2019). We test both a simple LSTM ("Model 3") (Hochreiter and Schmidhuber 1997), and a bi-directional LSTM ("Model 4") that leverages both the input sequence and a reversed copy in order to learn the whole context (Schuster and Paliwal 1997). Finally, we test transfer learning from two state-of-the-art language models, BERT (Devlin et al. 2018) and XLNET (Yang et al. 2019), using the Simple Transformers package (Rajapakse 2020). Transfer learning allows us to take advantage of pre-training on larger datasets and fine-tune a model to our particular classification task (Do and Ng 2005). The results of the 6 models tested are shown in Table 2.

On an imbalanced dataset, one can achieve high accuracy by simply always predicting the majority class. Rather, our goal is to identify as many recruitment-related posts as possible. Therefore, a predictive model that has many false negatives (recruitment posts that are predicted to be sales posts) is especially undesirable. Thus, we select Model 2 – which has the highest precision and recall on the validation set of all the models we tested – to be our predictive model class to use in the active learning process.

Model	Validation Precision	Validation Recall	Validation Accuracy
Model 1	89.3%	79.3%	92%
Model 2	**91.2%**	**82%**	**93.7%**
Model 3	88.6%	80.2%	94%
Model 4	83.8%	80%	93%
BERT	55%	72%	86%
XLNET	67%	65%	89%

Table 2 **Different DNN architectures tested prior to active learning process.**

Appendix B: Recruitment Templates

The active learning algorithm designed uncovered more than 27 types of recruitment tactics on the deep web.

Category	Definition
Adult Entertainment	Entertainment companies, bars, restaurants, strip clubs, bachelor parties, etc.
Escort	Agencies identified as escort services
Personal	Ads posted by individuals requesting personal interactions
Modeling	Agencies specifying jobs related to modeling
Porn	Ads recruiting for filming pornography
Massage	Ads recruiting for spas or massage parlors
Sugar	Ads recruiting for a sugar baby, a relationship where an individual provides money in exchange for an on-going relationship
Non-specified agency	Ads recruiting without specifying the type of work or job
Housing	Ads recruiting for vacant housing
Promotions	Job related to promoting products
Product Advertisement	Recruitment related to advertising products
Companionship	Ads specifying a paid companionship
House-keeping	Recruitment for house cleaning or cooking
Partnership	Ads recruiting for a business partner or escort partner
Make money	Ads specifying they can help you make money quickly
Walking	Recruitment for getting paid to walk
Booker	Recruitment for being a booker for an agency
Photography	Recruitment for exchanging photography for services
New Venture	Ads specifying partnering on a new venture
Finance	Recruitment for finance jobs
Club	Recruitment to join a specific club
Gangbang	Recruitment to be paid for a gang bang
Corporate Fitness	Corporate fitness jobs
Asian job	Roles specifying recruiting Asian women
Tourism	Recruitment for jobs related to hotels or tourism
Contest	Recruitment for contests
Videochat	Recruitment to get paid for a videochat

Table 3 Example recruitment templates identified across labeled posts

Appendix C: Auxiliary Results

We examine a number of relevant socioeconomic indicators (summarized in Table 4) to understand the characteristics of locations (in the United States) where vulnerable populations are deceptively recruited vs. sold for commercial sex. This data was collected at the county- or city-level across 8 government sources: US Census (Census 2018), US Bureau of Economic Analysis (BEA 2018), US Bureau of Labor Statistics (BLS 2018), US Department of Housing and Urban Development (HUD 2019), National Center for Education Statistics (NCES 2017), WomensShelters.org (Shelters 2021), Proximity One (Proximity 2009), and US Department of Justice (FBI 2016). The data collected from these sources focuses on both economic attributes (household income, GDP, unemployment) and social attributes (homelessness, education, crime).

We run separate Kolmogorov Smirnov tests (Smirnov 1939) to determine if there are systematic differences in the empirical distributions of each socioeconomic indicator in the top 50 'sender' versus top 50 'receiver' cities. We note that this is not a causal analysis since we are examining correlations. However, understanding the differences between recruitment and sales hubs can shed light on where different policy and social work interventions (e.g., those aimed at preventing victim recruitment vs. those aimed at rescuing current victims) would be the most impactful. Since we are testing a family of multiple related hypotheses, we employ the well-known Benjamini Hochberg procedure (Benjamini and Hochberg 1995) to maintain the resulting false discovery rate (FDR) at a standard choice of 10%.

We find that sender cities tend to be smaller (lower populations) and economically more constrained (higher poverty and lower household incomes). Sender cities also have more homeless people (i.e., vulnerable populations) and suffer high crime incidence (both property crimes and violent crimes). Figure 10 highlights significant differences in variables amongst sender and receiver cities. Together, these results suggest sender cities may not have as many resources as larger receiver cities to prevent trafficking of their vulnerable populations. Thus, operationalizing collaborations between counter-trafficking agencies along inferred recruitment-to-sales trafficking routes may significantly benefit resource-constrained sender cities in preventing victims from being trafficked in the first place.

Comparison to Rubmaps and Google Places

To the best of our knowledge, our study is the first to characterize recruitment in commercial sex supply chains, allowing us to uniquely identify trafficking recruitment risk at scale in commercial sex supply chains. In contrast, other empirical studies examine commercial sex activity purely from the sales side (e.g., through review websites such as Rubmaps), where the connection to human trafficking risk may be tenuous. We now examine how our deep web recruitment/sales densities compare to two such sources.

1. **Rubmaps:** Rubmaps.ch is a review site for massage parlors with sexual services, and has been used to assess commercial sex activity in prior work (Bouche and Crotty 2018, Diaz and Panangadan 2020). Rubmaps allows users to find and rate massage parlors by city/town. We manually extracted the count of massage parlors for each town listed on the website within the United States.

2. **Google Places:** Google Places includes a list of over 200 million global points of interest (e.g., restaurants, hotels, nail salons) that appear on Google Maps. We seek formally listed businesses with contact information (phone numbers or website) that also appear in the meta data of posts in our deep web dataset

Variable	Source	Kolmogorov Smirnov p-value	Statistical Significance after Benjamini Hochberg
Population	US Census (2018)	***	Yes
Real GDP	US Bureau of Economic Analysis (2018)	**	Yes
% of population with private health insurance	Proximity One (2009)	**	Yes
% of population with no health insurance	Proximity One (2009)	**	Yes
Violent crimes per 1000 people	US Department of Justice (2016)	**	Yes
Property crimes per 1000 people	US Department of Justice (2016)	**	Yes
Median household income	US Census (2018)	*	Yes
Poverty percent	US Census (2018)	*	Yes
Homeless per 1000 people	US Department of Housing and Urban Development (2019)	*	Yes
Homeless under 18 years old per 1000 people	US Department of Housing and Urban Development (2019)	*	Yes
Sheltered homeless per 1000 people	US Department of Housing and Urban Development (2019)	*	Yes
International migration per 1000 people	US Census (2018)	*	Yes
% of adults with bachelor's degree	US Census (2018)		No
% of adults with less than high school education	US Census (2018)		No
% of adults with high school education	US Census (2018)		No
% of students granted Pell Grants (federal subsidy for college)	National Center for Education Statistics (2017)		No
Women's shelters per 1000 people	WomensShelters.org		No
Unemployment rate	US Bureau of Labor Statistics (2018)		No

$^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 4 List of variables from governmental sources to compare the attributes of top recruitment cities against top sales cities in the United States. We employ the Benjamini Hochberg procedure to correct for multiple hypothesis testing.

from commercial sex advertisement websites; in other words, these businesses are likely associated with commercial sex sales, and therefore we refer to them as suspicious businesses. We find 5035 suspicious businesses, with 2630 listed in the United States/Canada. We manually categorize these suspicious businesses and find that the majority are spa/massage parlors (55%); other significant categories include home services (e.g., cleaning, repair, pool, roofing, moving), dollar general stores, and law firms.

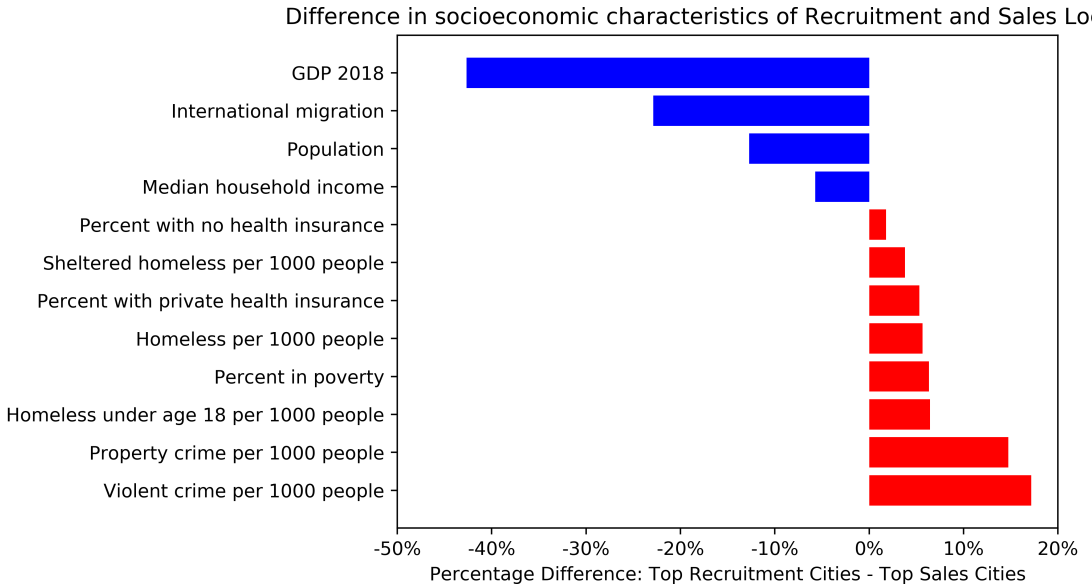


Figure 10 **Comparing selected socioeconomic variables for the top 50 sender (recruitment) and receiver (sales) cities in the United States. Blue and red bars indicate variables with higher values in receiver and sender cities respectively.**

We map these datasets based on city names to obtain heat maps of commercial sex activity (see Figure 11). Of the top 50 receiver locations we identified in the United States using deep web data, 82% included locations of suspicious businesses found in Google Places and 46% included massage parlors identified in Rubmaps; in contrast, of the top 50 sender locations, only 72% overlapped with suspicious businesses in Google Places and 26% with Rubmaps. Thus, we find that commercial sex sales activity identified on the deep web roughly aligns with activity identified through Rubmaps and suspicious formal businesses that may be selling commercial sex; however, recruitment activity is distinct and uniquely identified by our analysis.

Figure 11 **Empirical distribution of commercial sex activity in the United States inferred from Google Places, Rubmaps, and Deep Web respectively**