

Investigating Human Trafficking Recruitment Online: A Study Of Fraudulent Job Offers on Social Media Platforms

TOWERA JESSICA MOYO, University of Oxford, United Kingdom

OMER GUNES, University of Oxford, United Kingdom

MARINA DENISE JIRTOKA, University of Oxford, United Kingdom

In recent years, human traffickers have increasingly relied on social media to target and recruit victims. However, limited research has been conducted on the recruitment methods for human trafficking on social media, particularly in developing countries. Through in-person and digital observations, interviews, and case analysis, our paper sheds light on the tactics used by traffickers on social media to lure people into trafficking and highlights the characteristics of suspicious job advertisements used for human trafficking recruitment.

Our paper also discusses the investigative techniques employed by anti-trafficking experts to address recruitment and the challenges they face when identifying and investigating fraudulent jobs. Based on these findings, we provide recommendations on how technology can help tackle the problem of recruiting people online, and suggest design implications to improve the safety of social media platforms. Furthermore, we discuss the challenges and ethical considerations in developing anti-trafficking software and provide suggestions for addressing these challenges.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Human Trafficking, Social Media, Empirical Methods, Social Networking Site Design and Use, Empirical study that tells us about how people use a system, Ethnography, Qualitative Methods

ACM Reference Format:

Towera Jessica Moyo, Omer Gunes, and Marina Denise Jirtoka. 2025. Investigating Human Trafficking Recruitment Online: A Study Of Fraudulent Job Offers on Social Media Platforms. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW118 (April 2025), 31 pages. <https://doi.org/10.1145/3711016>

1 Introduction

Human trafficking is a global problem that affects millions of people around the world. More than 49 million people are estimated to be affected by this problem, with perpetrators making billions of dollars in profits by exploiting victims through activities such as forced labor and prostitution [9, 32, 50, 60, 109]. Human trafficking is mainly a real-world crime; however, in recent years, human traffickers have increasingly turned to social media platforms as tools to facilitate various forms of exploitation, such as sex work [9, 32, 50, 60, 109].

Human traffickers exploit social media platforms designed for communication, job searches, and social networks [12, 60, 86, 89] to recruit and exploit individuals, leading to dangerous situations [6, 13, 39]. These platforms have features that enable traffickers to easily access potential victims [36].

Authors' Contact Information: **Towera Jessica Moyo**, towera.moyo@cs.ox.ac.uk, University of Oxford, Oxford, United Kingdom; **Omer Gunes**, omer.gunes@cs.ox.ac.uk, University of Oxford, Oxford, United Kingdom; **Marina Denise Jirtoka**, marina.jirtoka@cs.ox.ac.uk, University of Oxford, Oxford, United Kingdom.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2025 Copyright held by the owner/author(s).

ACM 2573-0142/2025/4-ARTCSCW118

<https://doi.org/10.1145/3711016>

In 2018, the United Kingdom (UK) authorities discovered 539 pages on social media that promoted travel to Europe, taking advantage of people's desires in developing countries for a better life [20].

Recruitment represents a crucial stage in the human trafficking process; it is the first stage in which traffickers identify and lure victims. Unfortunately, research on the use of social media platforms to recruit victims of human trafficking is limited [47, 53, 59, 74, 79, 99]. Furthermore, most of the research on online recruitment has been conducted primarily in the Global North, particularly in the United States (US), leaving limited research on this topic in the Global South. Additional research on online recruitment would be beneficial in developing technical and non-technical solutions to address this problem, particularly in the Global South. By implementing preventive measures for recruitment, victims can be prevented from being exploited or subjected to traumatic experiences [108].

Anti-trafficking stakeholders, such as non-governmental and law enforcement organizations, work to identify instances of online human trafficking recruitment. However, there is also limited information on the methods used by these stakeholders to investigate and identify potential recruitment cases on social media. Obtaining such knowledge would be beneficial for developing tools to detect instances of human trafficking recruitment online and automate investigative processes to improve efficiency and scalability.

Two primary strategies are employed in online recruitment for human trafficking: posting fraudulent job advertisements or establishing fake relationships through grooming [9]. In the initial scenario, human traffickers employ fake job advertisements to attract potential victims, whereas in the latter scenario, they establish romantic or personal relationships. The investigation of recruitment through fake relationships presents particular challenges, as most interactions occur within private chat rooms, and access to such data is frequently restricted. As a result, this paper focused on fraudulent job advertisements.

In this paper, we will examine the online recruitment of victims in the Global South, with a focus on Africa. The study addresses three main research questions: 1) How do human traffickers use social media to recruit victims through fraudulent job offers? 2) How do anti-trafficking organizations investigate and respond to fraudulent job offers associated with human trafficking? 3) What are the typical characteristics of these fraudulent job offers in the context of human trafficking?

To address these research questions, we collaborated with an organization focused on identifying and investigating human trafficking recruitment cases in Africa. This organization operates in multiple countries and prevents people from being recruited online and offline. However, efforts to investigate fraudulent job posts for trafficking on social media are currently limited to one country, which serves as the focus of our research. The team in this country investigates one social media platform because most of the cases they handle originate there. Due to the organization's commitment to maintaining anonymity and preserving the integrity of their work, we will not disclose the names of the social media platform and the country in question. There are concerns that traffickers may alter their tactics on the platform, hindering the organization's ability to carry out their work effectively. Therefore, we refer to the country as "Country A", the organization in the focal country as "Organization A", and the platform of focus as "Platform A".

This paper makes the following contributions: 1) insights into the tactics employed by human traffickers who use fraudulent job offers as a means of online recruitment and the characteristics of these job advertisements; 2) an empirical understanding of the methods used by anti-trafficking organizations to investigate human trafficking recruitment on social media, as well as the challenges they face in their current setup; 3) insights into the challenges and ethical considerations of developing anti-trafficking software and recommendations to address these challenges; and 4) design implications for features that can be implemented on social media platforms to help reduce the likelihood of human trafficking recruitment.

Following the introduction section, the second section provides an overview of human trafficking and online recruitment, and the third section details the methodology used to address the research questions. The fourth section presents the findings, and the final section presents a synthesis of the results, including a discussion of the limitations and future directions of the research.

2 Background and Related Work

2.1 Human Trafficking

The United Nations Protocol on Trafficking in Persons [102], signed by member countries to address modern-day slavery, defines human trafficking as the act of recruiting, transporting, transferring, harboring or receiving individuals using coercion, abduction, fraud, deception, abuse of power or vulnerability, or the provision or receipt of payments or benefits to exploit them for financial gain. Human trafficking victims are exploited for various reasons, such as forced labor and prostitution, or other forms of sexual exploitation. Human trafficking in Africa is characterized by several factors, such as ineffective government leadership, political instability, economic challenges, and limited government capacity to combat trafficking effectively [37, 38, 44, 101].

Human trafficking typically involves three stages: recruitment, transportation, and exploitation. In the initial recruitment stage, traffickers employ various tactics to target vulnerable individuals, such as coercion, force, or fraud [71]. The transportation phase of human trafficking typically involves the movement of individuals between locations using various means of transportation, including land, sea, and air. However, it is important to note that human trafficking can occur without a transportation phase [105]. The final stage of human trafficking is the exploitation phase, in which victims are forced to engage in involuntary activities such as forced prostitution [71].

The digital age has led to innovative methods of conducting human trafficking, characterized by a shift from traditional physical interactions to dynamic online environments [36, 53, 74, 109]. This form of trafficking that occurs online is also known as internet-based trafficking, online trafficking, or technology-facilitated trafficking. Most of the research on online human trafficking has focused primarily on identifying patterns, keywords, and indicators related to how human traffickers advertise victims during the exploitation stage after they have been recruited [33, 49, 52, 60, 61, 114]. This body of work has left a significant gap in our understanding of online recruitment, which is the main focus of this paper. Previous research has called for more studies to understand how traffickers use social media platforms to target and recruit victims, including through fraudulent job opportunities [47, 53, 59]. This knowledge will assist stakeholders such as law enforcement and anti-trafficking organizations in developing measures to address this crime [53, 74, 99].

The following section discusses related work on online recruitment. The literature in the following section mainly discusses previous work from the Global North because most of the work has been done in this region, as previously mentioned.

2.2 Human Trafficking Recruitment on Online Platforms

In recent years, human traffickers have increasingly turned to online platforms for victim recruitment, making it one of the main methods used [30, 36, 39, 53, 74, 109]. The digital environment presents benefits to human traffickers, including the ability to reach victims in distant locations and the reduced probability of being detected by authorities [32, 60, 79, 109].

Victims are frequently recruited through the internet by unknown individuals who conceal their true identities [13]. It has been reported [118] that there are a substantial number of fake profiles on social media, allowing traffickers to anonymously search for victims through these profiles [13]. In addition, traffickers employ various techniques to maintain anonymity, such as concealing their

IP addresses, which poses a challenge to law enforcement and other stakeholders in identifying them, particularly their actual location [13, 104].

Traffickers use two primary strategies to recruit people online. The first approach involves traffickers targeting specific victims, typically on social media platforms, through messages that often begin as friendly, but quickly escalate to aggression as the relationship progresses [104]. On the other hand, the second strategy involves victims reaching out to traffickers in response to deceptive information posted online by traffickers [79, 104]. These victims are generally randomly selected based on those who react to this misleading online information, such as when fraudulent job advertisements are posted.

2.2.1 Fraudulent Job Advertisements. Traffickers use a variety of platforms, including social media marketplaces, reputable websites, and custom-built websites to lure people into trafficking situations through fraudulent job opportunities [79, 99, 104]. However, social media platforms are the preferred avenue for criminals to disseminate false job offers [39, 79, 95].

Fraudulent job advertisements often use attractive language to entice potential victims, promising better lifestyles [104]. Upon responding to these deceptive offers, individuals are subsequently exploited by traffickers who arrange their transportation. One such example is a case in which criminals established fake profiles on social media and advertised employment opportunities on various pages they created. They successfully recruited 100 young women through these pages and forced them to send compromised images [103]. It is common practice for traffickers to exercise control over their victims after they have been recruited and before they are transported [9, 79].

Volodko, Cockbain and Kleinberg [108] used indicators applied to identify labor trafficking in the real world to identify suspicious job advertisements on a Lithuanian website. However, it is necessary to develop indicators specifically for online fraudulent job advertisements for human trafficking rather than relying on indicators from the offline world [21]. Studies have concentrated on offline recruitment techniques and indicators, and only a few have examined online recruitment and its indicators, although it is a prevalent form of recruitment [39, 79, 104].

To develop effective indicators to detect fraudulent job advertisements on online recruitment platforms, it is crucial to gain a deeper understanding of the tactics employed by traffickers to create job postings to lure victims. Several studies have highlighted the need to investigate online recruitment, particularly how traffickers target and entice victims on social media platforms [47, 53, 59, 99]. Insufficient emphasis on online recruitment has resulted in ineffective strategies to address this problem [79, 119]. Understanding online recruitment will lead to more effective solutions to address this issue.

2.3 Anti-trafficking Policies and Online Legislations

Due to the numerous threats on social media platforms, various laws have been enacted to combat the abuse of these platforms and help protect individuals from crimes such as human trafficking. Most laws specifically for human trafficking usually target sex trafficking, because it is the most common form of human trafficking conducted online. These laws are summarized in Table 1.

The Communications Decency Act of 1996 drew criticism for facilitating human trafficking advertisements online because of the immunity afforded to Internet Service Providers (ISPs) and online platforms [25]. In response, the US introduced the SAVE Act in 2015, which held platforms responsible for trafficking content but with significant loopholes. To address these loopholes, the FOSTA-SESTA package was enacted in 2018, which amended the CDA to hold platforms responsible for sex trafficking content created by users [25, 40, 84].

In Europe, the Digital Service Act requires that online platforms in the EU address illegal content without monitoring user content, but by removing flagged content and using transparent automated

Table 1. Anti-trafficking Policies and Online Legislations

Policy	Impact on Anti-Trafficking Work
The Communications Decency Act (The US)	This act conferred immunity to online platforms regarding user-generated content, resulting in platforms hosting advertisements that promote human trafficking victims to potential clients without significant concerns for legal accountability [25].
The Stop Advertising Victims of Exploitation (The US)	Online platforms were held liable for user-generated content; however, this liability was removed if platforms attempted to moderate such content in any capacity, thereby introducing potential legal loopholes.
The Fight Online Sex Trafficking Act and Stop Enabling Sex Traffickers Act (The US)	Online platforms are held accountable for any sex trafficking content generated by users [25, 40, 84]. Online platforms demonstrated increased proactivity in removing suspected accounts; however, legitimate sex workers were inadvertently affected by these removals.
The Digital Service Act (The EU)	Necessitates online platforms to remove content about human trafficking identified by users or relevant stakeholders, such as law enforcement agencies, thereby prompting action against illicit material.
The Online Safety Bill (The UK)	It requires that platforms monitor content and remove any illicit material, such as identified human trafficking content, and implement measures to prevent access to such content while considering user privacy [46].
The Online Safety Act (Australia)	Mandates online platforms to assume responsibility for user safety by proactively safeguarding against access to illegal content, thereby compelling platforms to remove content associated with human trafficking [2, 72].
The Computer Misuse and Cybercrimes Act No. 5 (Kenya)	It facilitates the prevention, prosecution, detection, investigation, and punishment of cybercrime, thereby contributing to the mitigation of online human trafficking [69].
The Films and Publications Act (South Africa)	Statutory obligation for online content distributors and Internet Service Providers to safeguard the general public, with particular emphasis on minors, from content pertaining to human trafficking [43].
The Child Online Safety and Empowerment Policy (Africa)	This policy aims to protect children from online risks like human trafficking and to assist African nations in implementing legislation that seeks to ensure children's online safety [1, 4, 70].

moderation tools [35, 78]. However, the UK's Online Safety Bill and the Australia Online Safety Act of 2021 require platforms to ensure user safety by proactively monitoring and removing illegal content while considering user privacy [2, 46, 72]. All three regulations impose significant penalties for noncompliance.

In Africa, Kenya implemented the Computer Misuse and Cybercrimes Act No. 5 of 2018 [69] to help prevent, prosecute, detect, investigate, and punish cybercrimes. The Act also mandates the protection of accountability, integrity, and availability of computer systems, programs, and data. South Africa introduced the Films and Publications Act in 1996 and amended it in 2022 [43]. This

Act aims to ensure online safety by mitigating online harms, including provisions that require online distributors and ISPs to protect the general public from explicit and harmful content.

In 2024, the African Union Ministry introduced the first policy framework in the world for the implementation of children's rights in the digital space: the Child Online Safety and Empowerment Policy [1, 4, 70]. This policy aims to protect children from harms online and offline, which have become increasingly prevalent. This policy assesses the benefits and threats of digital platforms for children in Africa and identifies key issues and policy goals [1, 4, 70]. In addition, this policy aims to help African countries implement legislation that aims to protect children's safety online.

Despite various initiatives, efforts to combat cyber threats and ensure the safety of individuals online in Africa are disproportionately low compared to other continents [98]. Only a few countries have laws that address online harm, such as cyberbullying and human trafficking. As countries in Africa aim to establish legislation for digital platforms, policymakers would benefit from having access to more information on the role of the Internet in facilitating human trafficking in Africa to develop the necessary measures to address this issue.

3 Methodology

Due to the complex nature of human trafficking, investigating the topic necessitates direct engagement with anti-trafficking stakeholders to obtain meaningful insights. We devised a multifaceted research approach that involves a combination of in-person and digital observations, interviews, and case study analyses (see Figure 1).

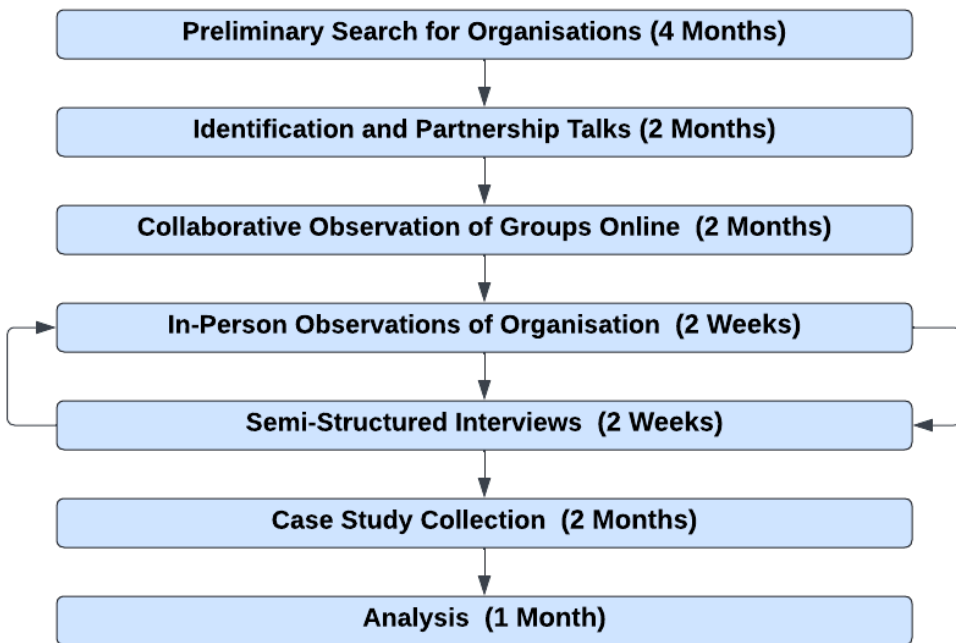


Fig. 1. Overview of the methodology process, from data collection to analysis

Each research method addressed at least one of the three research questions. The first research question, which focuses on the methods traffickers use to recruit victims, was primarily addressed through online observation of Platform A and supplemented by expert interviews to clarify the observed methods. The second research question, which examines how anti-trafficking organizations investigate and respond to fraudulent jobs, was primarily addressed through in-person observations of Organization A and interviews. The final research question, which focuses on the characteristics of fraudulent jobs, was addressed using all methods, including online observations on Platform A, case study analysis, interviews, and in-person observations. Incorporating in-person and online observations provides a more comprehensive approach to understanding research problems [68], which was one of the main reasons for including both methods to study the characteristics of fraudulent job advertisements.

The limited existing research in this area and the scarcity of publicly available datasets from anti-trafficking organizations make other data collection methods less feasible. In addition, obtaining data from social media platforms poses significant challenges, particularly for research on sensitive topics such as human trafficking. The following section outlines the process of identifying and establishing a partnership with Organization A.

3.1 Organization Identification

This research established collaboration with Organization A through the following steps:

- (1) Search engine queries with keywords such as "human trafficking organization Africa" and "human trafficking recruitment Africa" led to the identification of Organization A's website among the search results
- (2) Initial contact was made through an online form on their website, outlining the primary researcher's background, the research aims, and the need to obtain permission.
- (3) Organization A's administrator connected the research team with the head of data science, and initial meetings were held to gain a deeper understanding of Organization A's work.
- (4) The final steps involved obtaining ethical approval and simultaneously signing contractual agreements between Organization A and the research team's university, outlining how any shared data would be managed. Once these agreements were signed, the collaboration with Organization A began. Additional details about the ethical approval process are described in the following section.

3.2 Ethical Considerations

Ethical approval was obtained from the researchers' institution before starting the research. For online observations, a commitment was made to refrain from engaging with the participants. The ethical approval process clarified that deception would not be employed in research activities. During the study, participants' privacy and confidentiality were safeguarded by anonymizing all personal identification information from both online and in-person observations.

3.3 Participant and Platform Background

3.3.1 Organization A and Research Participants. Organization A is a local non-profit initiative that aims to prevent the recruitment of individuals into human trafficking in Country A. Organization A operates in a single city and maintains a physical office where its operations are headquartered. This initiative employs a multifaceted approach that seeks to educate potential victims about the dangers of human trafficking and prevent them from pursuing risky opportunities. Organization A investigates whether job advertisements on Platform A are related to human trafficking and prevents potential victims from pursuing these opportunities.

Organization A is made up of a diverse group, such as the project manager, the director, data analysts, and investigators. The investigators and project manager play a central role, responsible for identifying potential cases of human trafficking on Platform A by analyzing job opportunities. The lead researcher collaborated with five investigators, each with eight months to two years of experience, the project manager with seven years of experience, and the director with five years of experience.

3.3.2 Platform A. Platform A is a popular social media platform in Country A, with millions of users. It enables users to create profiles and connect with one another as well as establish online communities. Platform A offers various features, such as groups tailored to individuals with shared interests. This paper focuses primarily on job opportunity groups for Country A on this platform. Organization A centers its work on Platform A because of the significant number of suspicious job posts in the groups, the simplicity of conducting investigations, and the availability of relevant information, such as victim and trafficker contact details. Previous research has also linked Platform A to facilitating human trafficking in multiple countries [9, 32, 39, 60, 79, 104].

Job advertisement groups on Platform A vary in size from several hundreds to hundreds of thousands of members. These groups do not have membership prerequisites, making them accessible to individuals from diverse age groups and backgrounds. These groups can be categorized as public or private. Accessing private groups requires only an account on Platform A and a membership request. These groups allow for unrestricted posting and commenting by members.

The job advertisements analyzed in this study focused exclusively on the job advertisement groups associated with Country A. These groups predominantly included the name of the country or cities in their titles. In some cases, group names also included unique attributes specific to Country A, such as the name of the national high school certification. These groups are primarily targeted at residents of Country A, but can also be accessed by non-residents due to the ease of joining.

The job advertisements posted in the groups cover a wide range of occupational roles. The frequently advertised positions were that of domestic workers, housekeepers, security guards, and cleaners. These job advertisements are typically placed on behalf of companies or for personal reasons such as seeking assistance with household duties. Most job advertisement posts provide a description of the job in the post, while others include pictures with a job description or provide links to an external source containing further information about the job.

The job advertisements analyzed during this research met the following criteria:

- (1) The job advertisements were posted within one month of the review date. This time frame allowed Organization A to intervene promptly, as they also used the job advertisements in their work.
- (2) Job advertisements containing a company name and contact information. These details are essential for Organization A's investigations, with the rationale explained in more detail in the findings section.

3.4 Online Observation: Social Media Groups

We conducted an online observational study of 14 social media groups on Platform A over a two-month period, averaging 12 hours of active engagement per day. These groups were identified and provided by Organization A because they have been found to disseminate content related to human trafficking. The size of these groups ranged from 1500 to 143,000 members.

Before starting the data collection process, the researcher received training in identifying the indicators of fraudulent job advertisements related to human trafficking. Examples of such indicators include job advertisements that feature attractive salary packages and language patterns. Additional

details on indicators that can help identify fraudulent jobs for human trafficking can be found in Section 4.3. The primary researcher performed a manual review of job postings and their accompanying comments in chronological order in the 14 groups.

The primary researcher assumed a passive observer role, specifically observing group activities without engaging with participants who were unaware that they were being monitored [117]. Suspicious posts were saved in a private group on Platform A and forwarded to Organization A for further investigation. In total, 40 suspicious posts were reported to the project manager.

3.5 In-Person Observations and Semi-Structured Interviews

In-person observations were conducted by the primary researcher in Country A over an eight-day period within two weeks, with a daily commitment of four hours. The research participants were selected through purposive sampling, coordinated with the project manager. These observations followed an unstructured approach that involved direct interaction with the participants throughout this time.

During the observation period, notes were taken to record daily events and significant findings. The discussions with the participants were also recorded using audio devices during the observation sessions, and photographs of the environment and job advertisements were taken for the benefit of the broader research team.

During these in-person observations, semi-structured interviews [14] were conducted to supplement the information collected from the direct observations. These interviews, which lasted approximately 30 minutes on average, were designed to elicit the rationale behind the actions taken by investigators and managers and to clarify any observations made. The interviewees included the project manager, three investigators, and the organization's director, who were selected for their expertise and availability.

3.6 Case Notes Analysis

Organization A maintains a comprehensive database of all verified cases of human trafficking involving job advertisements. Each record in the database contains detailed information about the job advertisement, the investigation process, and indicators of human trafficking, as well as relevant information about potential victims, traffickers, and screenshots of the job advertisement. Organization A provided researchers with access to cases for four months, during which 15 cases meeting the scope of the study were analyzed to gain a deeper understanding and validate the insights obtained during the observation period.

3.7 Analysis

Data collected from audio recordings of observations and interviews were transcribed and entered into NVivo 12 for analysis. The lead researcher then used an inductive approach to develop codes by carefully examining observational notes from the ethnographic study, case notes, and transcripts from observations and interviews. The analysis produced a set of codes that were then developed into a codebook. The coding process focused on uncovering the investigative methods used by Organization A, recognizing patterns in suspicious job postings, understanding the tactics employed by traffickers on the platforms, and identifying any indicators that could help identify fraudulent job offers. NVivo 12 Pro was employed to code and draw patterns and conclusions from the data.

4 Findings

This section summarizes the research findings, with the first section addressing the results related to the first research question, the second section covering the second question, and the final section discussing the findings related to the third research question.

4.1 Methods Used to Recruit Victims on Social Media Platforms

Human traffickers predominantly use two primary methods to recruit victims from job advertisement groups on Platform A. The first involves the dissemination of fraudulent employment advertisements designed to elicit contact with potential victims. Victims typically respond by commenting on advertisements, providing contact information, or expressing their interest. Upon a victim's expression of interest, traffickers initiate follow-up communication through private messages on Platform A or via messaging applications such as WhatsApp. These job advertisements often generate rapid and substantial responses, with some receiving more than 200 comments in a few minutes. Many advertisements contain indicators of human trafficking.

In the second strategy, human traffickers monitor the content posted by potential victims within these groups to identify them. These posts often reflect individuals' job-seeking intentions, ranging from casual job inquiries to desperate pleas. Most of these urgent appeals come from people who have been unemployed and looking for work for an extended period. Traffickers exploit this by sharing their contact details, typically WhatsApp numbers, or by instructing people to initiate direct communication on Platform A. In some cases, potential victims include their own contact information in their posts, allowing traffickers to reach them easily. Individuals in these groups were observed to frequently publicly disclose their contact details.

These groups also attract individuals who promote financial schemes that promise quick and easy gains, often resulting in financial losses for potential victims. Many of these schemes exhibit indicators similar to those found in advertisements related to human trafficking, making it difficult to distinguish between them.

4.2 Overview of Organization's Workflow to Tackling Human Trafficking Recruitment

Figure 6 outlines the four main stages that Organization A employs to identify suspected instances of human trafficking and protect individuals from becoming victims of human trafficking. The initial stage involves monitoring groups on Platform A and selecting advertisements based on specific indicators. After the advertisements have been selected, Organization A conducts a thorough investigation to determine whether they are related to human trafficking. If sufficient evidence is found, Organization A takes the appropriate action by informing potential victims of the dangers of the job opportunity and discouraging them from proceeding. The following sections provide more details of these four steps.

4.2.1 Monitoring and Shortlisting Advertisements. The monitoring stage involves examining social media groups to identify job advertisements potentially linked to human trafficking. The investigators review postings from the predetermined list of 14 groups discussed in the Methodology section, along with other job advertisement groups within the country on Platform A. They occasionally use keyword searches to uncover suspicious job advertisements and then monitor related groups or pages. These methods are performed manually, without technological assistance.

Drawing on past experience, Organization A developed a set of indicators to identify fraudulent job advertisements in Country A. Investigators use these indicators to shortlist suspicious advertisements, relying not on a single criterion, but rather on a combination of various indicators to guide their selections. After identifying suspicious job advertisements, they employ various investigative processes to assess whether these advertisements are linked to human trafficking. These processes are discussed in the following section.

4.2.2 Investigation Process. Investigators typically operate individually when analyzing cases but occasionally seek advice from colleagues when aspects of a case are ambiguous or difficult to ascertain with high confidence. They frequently consult with the project manager for guidance

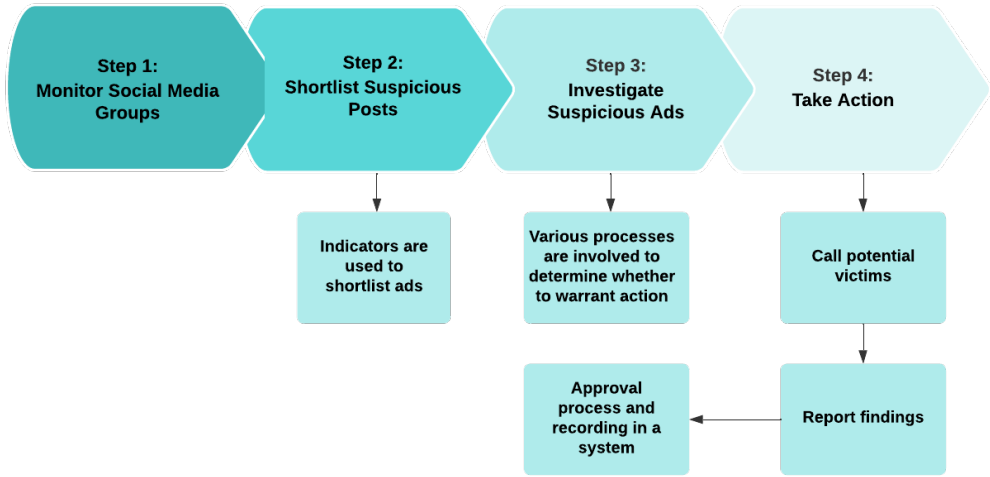


Fig. 2. Overview of Organization A's Workflow

because of his extensive experience in the field. The investigation procedure for each case can vary, but typically includes the following steps:

- (1) **Investigate the company** - The first step typically involves verifying the legitimacy of the company mentioned in a suspected fraudulent job advertisement. This process entails cross-referencing the company's information with a public online database containing records of registered companies, including their contact details, locations, and key personnel. Organization A verifies the authenticity of a company by searching its name in an online database. If the enquiry reveals that the company is non-existent or has ceased operations, additional investigative procedures are initiated. However, if the company is verified as legitimate, the next step involves comparing its details with those provided in the job advertisement. Any discrepancies or inconsistencies raise suspicions of human trafficking and warrant further investigation of the advertisement. In some cases, Organization A directly contacts the company to confirm whether they were the source of the job advertisement. Investigating the company is a crucial step, as it provides information necessary to establish a link between the job posting and human trafficking. The absence of a company name significantly hinders Organization A's ability to conduct thorough investigations and establish this connection.
- (2) **Investigate the phone number of the suspected trafficker** - Organization A employs an application to verify the validity of the telephone number associated with the suspected trafficker. This application provides information regarding the individuals connected to the phone number, including their name, registration status, and whether the number is spam. This application is publicly accessible and can be used by any individual. If there is a disparity between the name associated with the phone number and the information provided in the job advertisement, it raises suspicions about the legitimacy of the job post.
- (3) **Investigate the role and packages offered** - Organization A scrutinizes the offers of each job posting to validate its authenticity. By using multiple resources to ascertain the typical compensation package for the advertised position, Organization A can determine if the suggested remuneration is in line with current industry norms. The Project Manager states:

"One of the things that we encourage staff is to understand what is going on in the country and understand the different industries, the packages, and the requirements in the employment sector"

- (4) **Call the suspected trafficker** - Organization A contacts the suspected traffickers using the phone numbers provided in the advertisement and poses as a prospective candidate to gather information from them. During these conversations, they carefully observe any indications of deceptive behaviour, falsehoods, or exaggerated claims made by suspected traffickers. Investigators commonly pose deceptive inquiries to assess suspected traffickers' credibility, for instance, whether the absence of work records or an unlawful presence in the country impairs their capacity to pursue the opportunity. During the observation period, the researcher noted that each of the suspected traffickers contacted claimed that it would pose no problem if the investigators lacked legal authority to work in the country. On occasion, when suspected traffickers fail to provide their contact details in the job advertisement, the investigators leave a comment showing interest in the job on the post to facilitate communication. Typically, traffickers respond using messaging applications such as WhatsApp, allowing Organization A to obtain their phone number.
- (5) **Call potential victims** - Organization A seeks to gain more information by reaching out to potential victims and inquiring about any additional details they may possess about the job. Occasionally, this dialogue prompts victims to share more information, including specific details outlined in interview invitations, such as time and location. This additional information, such as the location, is then vetted for its potential association with human trafficking.

Organization A also employs these investigation techniques when potential victims reach out to verify opportunities. During awareness campaigns, they disseminate their contact details, encouraging people to verify the legitimacy of any opportunities they come across.

4.2.3 Taking Action and Reporting. Upon gathering a substantial amount of suspicion suggesting that the job may be fraudulent, Organization A takes a series of actions.

- (1) **Contact the potential victim** - To begin, investigators typically initiate communication with potential victims through telephone calls, during which they convey their suspicions about the job opportunity and provide explanations for their concerns. In addition, they advise against further involvement with the advertised job. Organization A does not engage in any further communication with the victim following this call. It is left to the victim to decide whether to continue with the opportunity.
- (2) **Write up the case** - Investigators carefully record all pertinent information related to the case, including the job advertisement, any interactions with the suspected trafficker, and any indicators that suggest human trafficking. This comprehensive case report is essential for future evaluation and authorization. At this point, the case is considered a "high risk of human trafficking".
- (3) **Approval process** - Following the completion of the writing stage, a data analyst verifies that the investigators have all the required case-related information before uploading them to a software platform. Once the case has been uploaded to the system, two senior personnel, one of whom is the project manager, conduct independent reviews of all relevant case information. Based on their assessment of the case, the two senior staff members each provide an individual recommendation on whether it should be classified as having "sufficient evidence of human trafficking" or simply as a "high risk of human trafficking".
When the two staff members do not reach a unanimous decision in their case assessment, a third senior staff member is consulted to make a final judgment. In most cases, disagreements

between the two members revolve around whether the case should be classified as high-risk or as having sufficient evidence of human trafficking. If deemed to have sufficient evidence, the case is officially documented as "evidence of human trafficking"; otherwise, further investigation is required.

- (4) **Record storage and access** - Following the approval of a case, it is no longer possible to make any changes to it. As a result, the case is recorded and archived on a digital platform used by Organization A to monitor and manage all cases it handles.

4.2.4 Challenges to Investigating Suspicious Advertisements. Organization A faces several obstacles when conducting investigations. The main challenges are lack of collaboration with external entities and scalability.

Collaboration. There is limited collaboration between organizations and law enforcement in Country A. Organization A works independently to prevent victims from being recruited, which hinders their ability to detain traffickers. Combating human trafficking requires collaboration between various stakeholders [28]. Organization A also confirmed that most anti-trafficking organizations in Country A face difficulties in working with various government agencies. This lack of collaboration between Organization A, law enforcement and the government has potential negative consequences for both staff and potential victims. For staff, working independently can be dangerous and can limit their access to information that could benefit their work. For potential victims, human traffickers can persist in their activities and continue to exploit new victims, thus perpetuating the cycle of human trafficking.

The director of Organization A, who oversees anti-trafficking efforts in 15 other countries in Sub-Saharan Africa, discusses two main reasons for the lack of collaboration in some countries compared to others with successful collaborations. Firstly, there are varying degrees of willingness among governments to work with anti-trafficking organizations. The Director states:

"It is different across countries. In some, there is a genuine willingness to collaborate with NGOs and provide additional resources. Some governments value the extra expertise and appreciate having dedicated support on the ground to combat human trafficking, which is a highly specialized crime. Often, their law enforcement lacks the capacity or depth of understanding to address human trafficking effectively." (Director)

Smaller countries also tend to work more with NGOs because it impacts their aid relationships with countries like the US, unlike larger countries such as Country A. The Director states that *"Smaller countries can really make quick wins in a way if they allow NGOs or they work with civil society, or if they just have the base levels of what they need to do, they can really move up that tier system quite easily"*. Governments draw on international frameworks such as the Palermo Protocol and the Trafficking Victims Protection Act (TVPA). The US Department of State evaluates countries based on their adherence to the minimum standards set forth in the TVPA, assigning scores that indicate the effectiveness of government efforts in tackling human trafficking [17, 106]. These evaluations affect the eligibility of countries to receive funding from the US.

The second reason for the lack of collaboration is that the state of anti-trafficking laws in different countries influences their willingness to work with other organizations. Many of these countries fail to comply with international best practices that encourage collaboration to combat human trafficking.

Previous research [73] confirms the lack of collaboration between governments and anti-trafficking organizations in Africa, attributed to various country-specific factors. Onuoha [73] highlights several aspects that raise concerns about the way African governments operate when fighting human trafficking, which could contribute to the lack of partnership. African leaders often prioritize their

interests over the allocation of resources and efforts to tackle human trafficking, with a greater concern for electoral politics. Furthermore, Onuoha [73] stated that corruption within law enforcement agencies also leads traffickers to evade justice. These issues highlight the much larger structural problems that affect successful collaborations.

Furthermore, in most African countries, government anti-trafficking strategies primarily emphasize prosecution, focusing on investigating cases, locating and rescuing victims, and apprehending traffickers, rather than on prevention [17]. In contrast, most anti-trafficking organizations, including Organization A, adopt a victim-centered approach that aims to prevent trafficking, identify victims, and provide assistance after rescue. This difference in objectives among anti-trafficking stakeholders can hinder effective collaboration.

Scalability. Organization A has limited financial and non-financial resources. The project manager stated that their mission is to make a significant impact with few resources. The manual method for identifying and investigating advertisements, combined with limited resources, presents a challenge in addressing the substantial number of fraudulent job posts. As a result, many fraudulent posts remain unattended, preventing Organization A from keeping up with the growing number of online recruitment schemes. Online recruitment has become popular because traffickers can target many more people than they would in the real world, thus expanding their criminal enterprise [36]. Organization A expressed concern that most of the employment opportunities reviewed within these groups are fraudulent.

In addition, many positions are not evaluated due to the criteria used by Organization A in their examination process. They have a policy to evaluate job advertisements only if they include contact information for both the trafficker and at least one potential victim. This leads to many suspicious job postings not being investigated. A call to the trafficker is a crucial aspect of the investigation, and their ability to intervene is limited to preventing the individual from pursuing the opportunity, as Organization A does not collaborate with law enforcement agencies. Contact information is obtained from the job advertisement and the corresponding comments section. However, traffickers and victims do not always openly share their contact information, often instead relying on private messaging.

4.3 Identifying Suspicious Job Advertisements

The following sections describe the characteristics of job advertisements used to determine their potential for human trafficking. The educational background and experiences of the victims can sometimes hinder their ability to recognize certain indicators outlined in this section.

To maintain anonymity, the images of the job advertisements were modified by removing timestamps and other identifying graphics using cropping tools. These indicators are grouped into five main themes that highlight the characteristics of suspicious job advertisements. Table 2 presents an overview of the themes and their associated indicators.

4.3.1 Attract Wide Audiences. Suspicious job advertisements often employ a strategy of casting a wide net to draw in many candidates, thus enhancing their chances of successful recruitment. This is typically achieved by using four main methods: the use of *well-known company names*, the advertisement of *multiple roles and mass recruitment*, and the inclusion of *specific preferences*.

Company Names. A job advertisement found on Platform A's groups raises concern when it includes a role from a *reputable company*. Established entities typically use more formal recruitment channels than social media. Traffickers who are suspected of using such channels often select well-known companies, particularly those in the food industry. Organization A has observed and confirmed on multiple occasions that recognized companies rarely resort to Platform A for their

Table 2. Themes and Indicators

Theme	Indicators
Attract Wide Audiences	Company Names Recruitment Volume Multiple Roles Preferences
Use of Informal Writing	Writing Styles Emojis Suspicious Links
Offer Attractive Packages	Salary Incentives
Use of Known Trafficking Details	Similar Job Advertisement Details Trafficking Locations
Request for Personal Information	Personal Details

hiring needs. Dhaliwa [31] has also stated that legitimate companies do not use social media platforms known to connect people for personal reasons to advertise job opportunities.

Recruitment Volume. Human traffickers commonly publish job advertisements that claim to recruit *substantial numbers* of individuals, sometimes even hundreds. This trend raises suspicions about scam activities, as it is rare for companies to recruit a large number of candidates in a single post or all at once. For example, Figure 3 illustrates a typical job advertisement that recruits many people. In particular, the format of these posts differs significantly from the formal style typically employed by legitimate companies in their recruitment efforts. A potential individual was on the verge of falling victim to a similar offer before Organization A intervened to prevent them from accepting the opportunity.

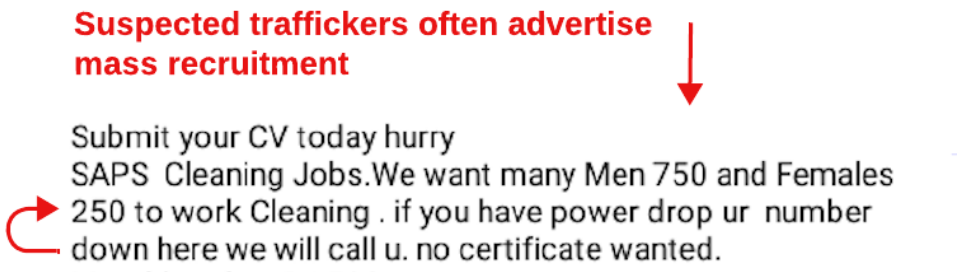


Fig. 3. Indicator showing mass recruitment

Multiple Roles. Suspicious job posts frequently advertise multiple roles within a single listing, which can appeal to a wide range of applicants by offering a variety of employment opportunities.

These advertisements warrant scrutiny, especially when they list more than three positions, as they often depart from the standard format of legitimate job posts. In addition, the advertised roles are typically those that pose a high risk of human trafficking. Examples of the roles are security guards, general workers, cleaners, and cashiers.

Preferences. Job advertisements that express *specific preferences*, such as demographics, should be viewed with caution, as they can be exploited by traffickers to expand their reach and target vulnerable individuals. For example, traffickers can remove or minimize job qualification requirements to attract people without experience or specific qualifications, thus increasing the number of potential victims. In addition, traffickers can target vulnerable individuals by specifying countries with a high number of undocumented immigrants, as these individuals are often desperate for job opportunities. Such job postings may be advantageous for traffickers, as they can entice many potential victims with the promise of employment. The job advertisement illustrated in Figure 4 indicates that no previous experience was required to apply, which encouraged potential victims to express their interests and contact the trafficker. Fortunately, Organization A was able to verify its fraudulent nature and prevent one of the victims from being recruited. "No experience required" is a common phrase that can be searched to identify suspicious content. The *writing style* of these posts also serves as a critical indicator.

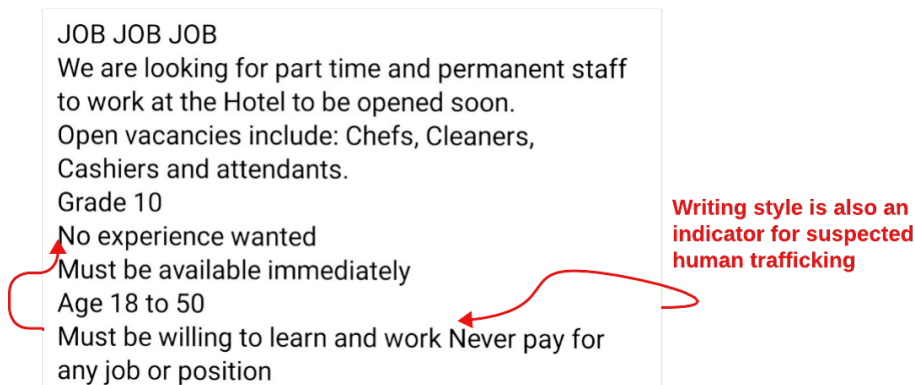


Fig. 4. Indicator showing writing style

4.3.2 Use of Informal Writing Styles. Advertisements of suspicious nature typically employ an informal tone, including *informal writing*, *emojis*, and *hyperlinks*. These elements are not typically found in genuine posts and can be indicative of an attempt to deceive.

Informal Writing. The *writing style* is also a significant indicator to identify suspicious posts. Traffickers often create job postings with grammar and spelling errors as well as casual language, in contrast to the formal tone typically found in legitimate company advertisements. In this writing style, the traffickers aim to entice people to consider the opportunities. For example, as shown in Figure 4, job advertisements contain grammatical errors and cause concern.

Emojis. The use of *emojis* in job advertisements is not a common practice and their presence often raises concerns. Traffickers may use multiple emojis, such as symbols of money or flags, to capture the attention of potential victims and lure them into applying for a job. People seeking

employment in developing countries are commonly looking for opportunities in the Global North. Thus, emojis that represent money or countries can quickly gain their attention. An example of this is shown in Figure 5, which shows a job advertisement with Canadian flags.

Links. Traffickers often include misleading links in their posts, which is another indication of fraudulent job advertisements. These links typically lead users to fake job pages rather than promised job opportunities. The existence of such deceptive link patterns is considered when assessing whether a post is associated with trafficking. In addition, attractive packages are also a significant indicator.

4.3.3 Offer Attractive Packages. Traffickers typically format their advertisements in an appealing way to attract potential victims and entice them to apply for opportunities. This section explores *salaries* and various *incentives*.

Salary. Traffickers often resort to offering unusually *high salaries* as an incentive to recruit victims. Such salaries are often considerably higher than the average remuneration for comparable positions in a country. Experts familiar with the prevailing salaries and job packages in a given nation can easily discern these unrealistic offers.

Incentives. Human traffickers often advertise job openings with attractive incentives, such as free transportation and housing, to entice people to respond to their advertisements. There have been several cases in which Organization A has encountered people who were offered free transportation for interviews. Using these attractive offers, traffickers can capture the attention of potential victims. In Figure 5, a trafficker promised to cover the costs of visas and flights for potential victims, leading several individuals to contact the suspected trafficker. Organization A has indicated that traffickers often assume these expenses when transporting them to another country, but subsequently use them to establish control over victims through debt bondage if the victims are unable to repay the costs.

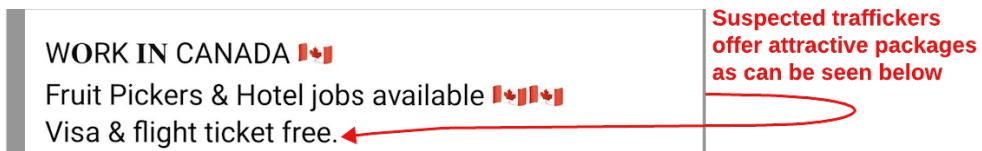


Fig. 5. Example of Job Advertisements - Indicator offering attractive packages

4.3.4 Use of Known Trafficking Details. Identifying suspicious job advertisements can also be achieved by uncovering specific details associated with known offenders from previous cases. This section explores the importance of *similar advertisements* and *well-known locations* as important indicators in this regard.

Similar Job Advertisements. Traffickers frequently resort to duplicating information when creating deceptive advertisements. Organization A can identify such advertisements by referring to previous investigations. Suspected traffickers typically modify job postings marginally before republishing them on Platform A. If a job advertisement contains personal information or content that closely aligns with a previously confirmed case, the advertisement is flagged. In addition,

individuals engaged in trafficking activities often utilize the same account to post new jobs. The project manager comments on this issue and states the following:

"These people use the same name for everything. It is not the first or second company that I have seen with the same name. You know, they might use one phone number for WhatsApp communications" (Project Manager)

Trafficking Locations. The presence of *known trafficking locations* in Country A plays a pivotal role in identifying suspicious job advertisements. Whenever a job listing mentions these well-known trafficking spots, Organization A flags the post. Occasionally, community members leave cautionary remarks on such posts to deter individuals from falling prey to such schemes.

4.3.5 Request for Personal Information. Job advertisements that solicit personal information in the comments section warrant scrutiny, as human traffickers frequently request the contact details of potential victims to initiate private communication. Traffickers often employ these details to establish communication with people about interview invitations. In some cases, traffickers gather contact details about potential victims from other posts within the group and use them to send them interview invitations. Individuals often express skepticism about job offers within these groups, especially when they receive interview invitations from companies they never applied to, which raises concerns.

4.3.6 Summary and Account Profiles. The five themes discussed above can be summarized into three main categories that investigators aim to detect: 1) incongruities, 2) irregularities, and 3) familiar misdemeanors. Investigators are continuously looking for incongruities, which are elements of information that contradict or deviate from industry standards, and a broader understanding of the national job market, such as advertisements that attract wide audiences. Organization A ensures that researchers receive regular training on the contextual aspects of the job market, whilst also mandating them to undertake independent research.

Identifying oddities requires detecting information that deviates from norms, such as unconventional practices or details that stand out as atypical. These norms may include the use of casual language styles and the improper solicitation of personal data. To detect such oddities, Organization A meticulously examines job postings and scrutinizes each line to uncover subtle elements that deviate from standard recruitment advertisement practices.

Investigators also rely on identifying familiar misdemeanors by detecting similar information posted by traffickers. They achieve this by keeping a record of information associated with trafficking, collecting data on confirmed details related to human trafficking, and reviewing new information against these records.

Suspicious advertisements can also be identified by analyzing account profiles within the job advertisement groups. Several patterns were discovered in the accounts of suspected traffickers, which can help identify them and their job advertisements. These account profiles exhibit various characteristics that can be flagged:

- **Account profile** - Individuals who posted suspicious advertisements on social media typically had profiles containing only job-related content. On the platform, users can view the previous posts of an account holder if their profile is public. Most of these account holders' content consisted of job posts in a variety of industries and companies, with many exhibiting suspicious characteristics.
- **Account Photo** - The account profile's photos were also a point of suspicion, as many of the individuals who posted dubious content did not have a profile picture or had a picture that did not depict themselves. Instead, their photos contained quotes or other elements. In some instances, the accounts feature photographs of well-known models.

- **Connections** - The accounts under scrutiny typically exhibited a lack of connections or a small number of connections, raising suspicions and necessitating further investigation. This suggests that the account is predominantly employed for the dissemination of content about job opportunities.

These indicators of fraudulent job advertisements and account profiles can help various stakeholders combat human trafficking.

5 Discussion and Recommendations

This section discusses the findings and highlights opportunities for using technology to combat human trafficking, the challenges of developing detection software, and the ethical considerations involved. It also offers design recommendations for social media platforms to help prevent human trafficking and presents a framework for developing and deploying anti-trafficking software.

5.1 Opportunities to Tackle Online Recruitment Using Technology

The findings highlight the opportunity to use technology to help stakeholders combat fraudulent jobs for human trafficking recruitment online. This section focuses on the various ways that technology can be used to address this problem.

5.1.1 Reviewing Job Posts Online. This paper highlight the need for a tool that can identify suspicious posts related to human trafficking. Automating the process of identifying and shortlisting suspicious posts can significantly enhance Organization A's efficiency in handling a greater number of job advertisements, while also enabling Organization A to reallocate resources to other critical tasks. This aligns with previous literature [92], which suggests the development of a tool to detect suspicious job posts online. Current anti-trafficking software focuses mainly on the exploitation stage when victims have already been recruited [8, 9, 32, 33, 41, 52, 59–61, 86, 90, 92, 99, 100, 111, 112, 115, 120]. However, there is a lack of tools that target the human trafficking recruitment stage despite evidence that most recruitment activities occur online in several countries such as the US and Italy [39, 79, 104].

Software aimed at combating trafficking during the recruitment stage could benefit from integrating advanced technologies, such as a large language model (LLM) driven intelligent system, to analyze and classify suspicious job postings. LLMs have been shown to outperform traditional machine learning approaches and also excel at classifying social media posts [22]. LLMs present a valuable opportunity in this field, particularly where access to labeled data is limited, as they deliver high performance with relatively small datasets. In contrast, traditional ML models typically require extensive training data [18, 62, 81].

5.1.2 Investigating Job Posts. Organization A, as previously mentioned, conducts various investigative procedures following the shortlisting of suspicious advertisements, and on occasion, potential victims approach Organization A to confirm job opportunities. Developing a tool that enables people to request verification of employment opportunities from Organization A would streamline the process, facilitating the review and management of cases. This approach would be an improvement to the current practice of using a messaging application, such as WhatsApp, to make contact with Organization A. Furthermore, this tool would allow Organization A to initiate investigation procedures based on these requests or upon identifying suspicious job posts online.

This software tool could include features for investigating jobs such as verifying the legitimacy of companies, phone numbers, and locations in job advertisements. For example, the tool could incorporate a company checker to investigate various aspects of a company's information, such as whether a company mentioned in an advertisement exists, its operational status, and contact details.

This tool would enhance Organization A's efficiency in performing its duties and investigating more cases, thereby increasing the scalability of cases handled.

5.1.3 Identifying Potential Traffickers and Organized Networks. The findings indicate the existence of patterns in the profiles of suspected traffickers. Organization A also noted that traffickers often reuse job posts from the same or different accounts. Social media platforms have the opportunity to flag accounts that repeatedly post the same suspicious job postings. They can also identify potential organized networks by flagging accounts that share suspicious job posts with identical details, such as location and contact information.

Previous research has also shown that there are networks of traffickers operating on various online platforms that exploit multiple victims [10, 11]. In this case, a network of traffickers means a group of traffickers operating online to carry out illicit activities. These groups can range from a small number of traffickers exploiting one or a couple of victims to large networks involving influential figures operating in various countries [10]. There are opportunities to use various technologies to flag potential traffickers and organized trafficker networks on social media. Trafficker networks can be identified on the basis of their connections and interactions. Ardakani [10] conducted a study to identify organized sex trafficking groups in Louisiana using online advertisements. The authors connected advertisements using advertisement IDs, phone numbers, and textual content of the advertisements using author attribution techniques. In contrast, Arenas et al. [11] identified trafficker networks that promote OnlyFans accounts of their victims on Twitter. Both studies demonstrated the benefits of social network analysis in identifying trafficker networks [10, 11].

Limited research has used content generated during the recruitment stage to identify potential traffickers and their networks. Identifying organized networks during this stage can help platforms flag traffickers and prevent them from sharing advertising opportunities, thereby reducing the recruitment of individuals. Furthermore, there is a potential opportunity to verify whether organized behavior can be identified through mentions and links in job posts, similar to the approach taken by Arenas et al. [11] in analyzing OnlyFans content.

Implementing these recommendations has significant potential to help combat human trafficking recruitment on online platforms. However, several challenges may hinder the performance of these tools, potentially affecting their effectiveness in both the short- and long-term. These challenges are discussed in the next section.

5.2 Challenges of Using Automated Methods to Identify and Investigate Recruitment

There are potential obstacles to effectively developing anti-trafficking tools or features for human trafficking recruitment, whether for monitoring or investigation.

Accuracy. Achieving accuracy in anti-trafficking software is challenging due to the lack of ground truth data. Most solutions rely on expert-labeled data, which can be prone to errors, affecting the accuracy of the model and the trust in these tools [92, 100]. Furthermore, most of the posts reviewed in this paper showed a tendency to use informal language, complicating accurate classification, since slang and vocabulary vary by country. The informal text on social media can significantly influence the accuracy of machine learning models [7, 83]. Limited data availability further hinders model development, as organizations are often reluctant to share data due to privacy and proprietary concerns.

Compliance. Meeting regulatory requirements is another significant challenge. Compliance with data collection and processing laws, such as the General Data Protection Regulation (GDPR), affects tool design and deployment feasibility [21]. For example, these laws can restrict the collection and processing of data obtained through automated means such as web crawling, hindering the

analysis of information left by traffickers and victims. Most data related to human trafficking for automation purposes is derived from crawling the web for advertisements written by traffickers, such as those that advertise victims on classified sites and job boards.

Integration. Entities like Organization A require access to various types of data to conduct investigations, which poses a challenge when developing tools to aid the investigation process. Such tools would need collaboration with third-party applications. However, collaboration between anti-trafficking stakeholders is challenging [21], as public and private organizations are often hesitant to share data or integrate their modules. An alternative is for stakeholders to create their own datasets, such as a company database to verify companies. However, this approach may not be feasible for all needed data such as phone numbers because of privacy concerns.

Long-term Performance. The long-term performance of anti-trafficking software can be affected by factors such as feedback loops and concept drift. Feedback loops, in which an AI system's predictions influence future data, can reinforce biases in AI systems, leading to a narrow focus on specific cases [93, 113]. Concept drift occurs when statistical properties change over time, making historical patterns inapplicable to new data [42, 54, 87]. Human traffickers tend to adapt their methods to avoid detection, such as using emojis instead of text, increasing the likelihood of concept drift, degrading the performance of the model [114]. Concept drift and risky feedback highlight the importance of developing adaptive AI systems that can continuously learn and evolve to effectively counter new tactics and maintain accuracy in identifying trafficking activities.

Feasibility. Anti-trafficking organizations and other stakeholders typically operate with limited resources. Developing, implementing, and maintaining anti-trafficking tools requires significant resources [21]. In addition, meeting the growing requirements for detecting trafficking activities can be financially challenging. Organization A and previous research [30] have confirmed that traffickers use several social media platforms and job boards to expand their operations. Limited financial and non-financial resources impact the ability to scale operations, particularly in resource-constrained regions like the Global South.

In addition to the challenges discussed in this section, there are also ethical concerns related to anti-trafficking tools, which are detailed in the following section.

5.3 Ethical and Privacy Considerations for Developing and Deploying Anti-Trafficking Software

A major concern when using anti-trafficking tools is the presence of data bias [48, 97]. Data bias can arise from various sources, such as a lack of ground truth data, existing biases in training data, and implicit human biases during tool design [75]. Data annotation is a significant contributor to data bias and is influenced by factors such as race and gender [51, 67, 75]. Biases in anti-trafficking software can lead to false positives, negatively affecting marginalized groups.

Misidentification of individuals as traffickers or victims can have significant consequences, such as stigmatization of those incorrectly identified as traffickers and denial of access to certain resources for misidentified victims (Author Work Under Review). For example, some anti-trafficking software has been associated with adverse outcomes for individuals using online platforms in the US. Following the implementation of the FOSTA-SESTA Act, several platforms, including Craigslist and Reddit, removed all sections, such as forums related to sex work, and now employ AI models to block the content associated with sexual activity, resulting in the ban of sex workers [23, 63, 91]. This loss of digital space forced many sex workers to operate on the streets, placing them in dangerous situations and increasing their vulnerability to violence and exploitation [23]. A sex worker reported that within the first month of the law enactment, 13 people were reported missing

and two were confirmed to be deceased [91]. The SAFE SEX Workers Study Act was introduced [80] to address the unintended consequences of FOSTA-SESTA. This legislation aims to investigate the effects of FOSTA-SESTA on sex workers and use the findings to inform potential amendments or possibly even lead to the repeal of the law.

The implementation of anti-trafficking tools that target recruitment may also lead to unintended consequences. For example, online platforms may erroneously ban legitimate companies because of false associations with fraudulent human trafficking job advertisements. Moreover, job advertisement groups can potentially be removed in a manner similar to the sex work forums discussed above. This could adversely affect individuals who post legitimate content and impede job seekers' ability to find employment opportunities if the groups are completely eliminated.

In addition, developing a software tool that analyzes job listings on a platform raises privacy concerns due to the sensitivity of the data it would access. Such tools would have access to posts that may contain personal information, such as contact details, thus potentially exposing them to harm. Organization A also requires personal information for their investigations, and if a tool collects this information, it could pose privacy risks. Previous research has highlighted concerns about potential privacy breaches associated with the use of these types of anti-trafficking tools [27, 59, 96]. Tools may inadvertently collect data that are not related to human trafficking, such as data on voluntary sex workers [27]. Therefore, there is a concern that information collected for investigating fraudulent job advertisements could also be used for purposes other than job classification. Anti-trafficking software has the potential to be misused for other negative purposes, such as illegally tracking individuals or companies [28]. Any errors or missteps in this area can have severe consequences [28, 74], potentially causing harm to victims and other individuals in society. Therefore, it is important to develop strategies to preserve privacy and protect important information during the design stage.

5.4 Design Implications for Social Media Platforms

Social media platforms are obligated to protect users from harm online, including human trafficking [3]. Social media platforms should take a proactive approach to combating human trafficking recruitment. We present a set of recommendations for social media companies to consider when designing, developing, and iterating their platforms to promote safety against human trafficking recruitment.

5.4.1 Stringent Guidelines. We recommend that designers implement an additional layer of security by imposing stricter criteria for the entry and posting of content in job marketplaces and groups. The current process of admitting individuals to groups poses a risk to vulnerable individuals. Social media platforms should proactively establish and enforce rules for job opportunity-related groups, rather than rely solely on community moderation and group administrators. For example, social media platforms possess the resources and data necessary to develop effective methods to detect and remove fake profiles from their networks and prevent them from accessing job marketplaces and groups. In the last half of 2023, LinkedIn used automated and manual processes to stop and remove more than 60 million fake accounts from their platform. Furthermore, Facebook sanctioned more than 630 million accounts in the first quarter of 2024 [66]. The prevalence of fraudulent profiles on social media has allowed traffickers to easily identify victims [13]. Social media platforms can also use the additional indicators identified in this study to flag suspicious accounts attempting to join job advertisement groups, alongside their efforts to remove fake accounts.

5.4.2 Warning Mechanisms. Social media platforms should integrate warning labels for users who join job marketplaces and groups to inform them of the potential risks associated with such

communities. By raising awareness of human trafficking and other illegal activities, people can exercise greater caution when seeking online opportunities. Platforms can also send reminders after a certain period to remind users of the dangers associated with these communities. Research has shown that occasional reminders can have a positive impact on behavior in various areas, including health [77]. In addition, warning labels for online misinformation have been shown to be effective in increasing user awareness of misleading content [24, 64].

5.4.3 Cautionary Messages. Social media platforms should implement cautionary messages to discourage users from sharing personal information such as their contact details, as traffickers often exploit this information to lure victims to participate in fraudulent interviews. By displaying informative notifications that highlight the potential risks of sharing such information, such as human trafficking, users may be better equipped to avoid falling victim to recruitment scams. Furthermore, social media platforms can flag job advertisements that ask users to publicly disclose their contact details because genuine job listings generally do not require this information to be shared openly. However, it is important to consider the potential unintended consequences of such measures, as users may be annoyed by constant reminders and warnings. Therefore, social media platforms should explore less intrusive options or allow users to turn off these notifications if desired.

5.4.4 Moderation. Social media platforms should take a proactive approach to analyzing job posts and account profiles for signs of human trafficking to prevent them from posting job advertisements. Platforms can use human and automated moderation. Automated moderation can involve mechanisms such as time delay can be implemented while a post is reviewed. Content moderation should always involve experienced individuals in making final decisions before implementing drastic measures, such as removing an account from the platform.

When anti-trafficking software is developed, it is crucial to consider several factors to ensure that ethical concerns and challenges are adequately addressed. The following section explores some of these key factors.

5.5 A Framework for Developing and Deploying Anti-trafficking Tools

This section presents key factors that stakeholders should consider when developing and deploying anti-trafficking tools.

5.5.1 Stakeholder Collaboration. The development process should engage with various relevant stakeholders during the design stage, especially anti-trafficking organizations and law enforcement, who are some of the main users of these types of tools [74] and are also some of the most knowledgeable about the crime. To help identify stakeholders in this field, especially law enforcement, researchers can use mailing lists such as Rephrain [82], which includes various stakeholders working on online harms, including law enforcement. In addition, using professional platforms such as LinkedIn to conduct customized keyword searches can help identify people with relevant expertise.

Similarly to this research, anti-trafficking organizations can also be identified using search engine queries. A few observations were made during the collaboration process:

- (1) Efforts should be made to find common areas of interest to provide incentives for anti-trafficking organizations to partner with academic researchers.
- (2) Researchers should aim to write a clear message that highlights their background and research plan to increase credibility, while ensuring that the incentives are evident to prompt responses.
- (3) Efforts should be made to start discussions with institutions early regarding contractual agreements, such as NDAs, as these can lead to delays. The agreements should ensure that

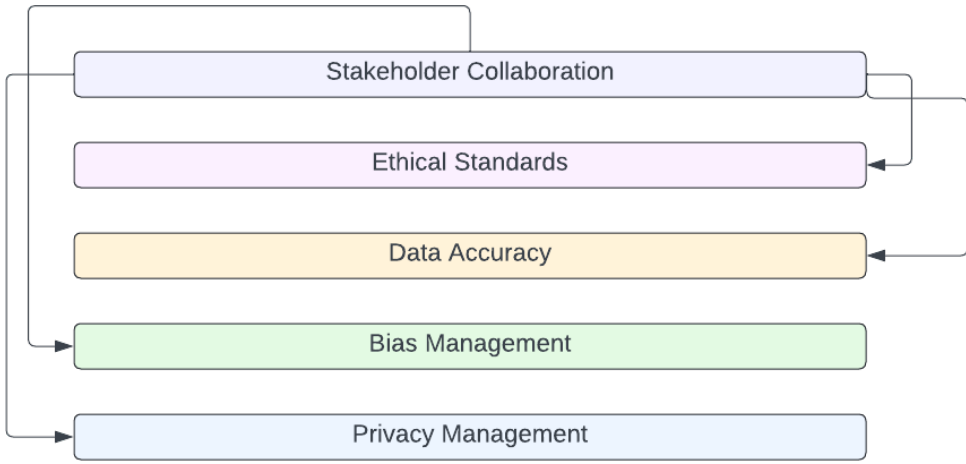


Fig. 6. Framework for developing and deploying anti-trafficking tools

the integrity of the organization is maintained, while allowing academic researchers the freedom to publish and present results.

5.5.2 Ethical Standards. Human trafficking is a sensitive subject that requires careful consideration. There is a need to develop best ethical guidelines and practices for researching online trafficking and developing anti-trafficking software. These practices and guidelines should be developed closely with experts in the domain. An ethical review process and risk assessment should always be conducted before beginning a project, involving domain and non-domain specialists, to assess the project's plans and strategies and whether they meet the ethical standards. The ethical evaluation will allow stakeholders to make well-informed decisions and determine whether the proposed project complies with ethical standards throughout the development and post-deployment stages. The risk assessment will also help identify any risks in addition to privacy and bias that can occur with technology and ways to mitigate such risks.

5.5.3 Data Accuracy. Data annotation is crucial in anti-trafficking tools due to limited ground truth data. Stakeholders should involve experts in annotating unlabeled data to ensure that domain knowledge is accurately represented. At least one expert should participate in an iterative labeling process until the data quality is satisfactory. Involving multiple experts allows developers to validate labels using methods such as the Inter-Annotator Agreement (IAA), which measures the agreement between annotators [19, 116].

After creating or gaining access to a dataset, it is necessary to ensure its high quality to maintain accuracy and comply with industry and government regulations [34]. We recommend using both quantitative and qualitative evaluation methods before training the model. Automated inspection can detect and rectify problems such as spurious correlations [48, 85, 94, 110], enhancing data accuracy and consistency while minimizing human errors [34]. Despite the efficiency of automated dataset assessments in identifying certain issues, human-conducted inspections can be advantageous in revealing other problems that may not be readily apparent, such as biases related to race and gender. Furthermore, manual audits can offer valuable insight in methods to improve data quality

[57]. Kreutzer et al. [57] state that human-led inspections are particularly effective at recognizing subtle errors, inconsistencies, and other issues that could evade detection by automated methods alone. This is very important for a field like human trafficking where experts are encouraged to be in the process of developing the tool.

5.5.4 Bias Management. Stakeholders should develop improved data collection and annotation methods that reduce instances of data bias [75, 88] by ensuring diversity in the data [88]. We also recommend that the teams responsible for collecting and annotating data should be diverse and consist of representatives from various demographics and skill sets. Stakeholders should include members from marginalized groups that are affected by these systems, such as victims [29].

Furthermore, stakeholders should consider employing algorithms and tools to eradicate bias in datasets using debiasing strategies. These algorithms and tools, such as DRiFt (Debias by Residual Fitting) [16], apply statistical techniques to identify and mitigate biases [15, 16, 48, 65]. Machine learning models have shown poor performance for people of color and females [65], which should be considered when developing anti-trafficking solutions. Most automated data bias correction methods involve the use of statistical techniques, which raise concerns about potential inaccuracies; therefore, it is important to combine them with expert validation. Furthermore, it is crucial to assess the consequences of the dataset when using algorithms because some deletion processes can lead to additional issues, lower accuracy, and impact the decision-making process of a model [56].

5.5.5 Privacy Management. Stakeholders must be transparent about their plans to protect user data from the start of the project. To address privacy risks, several aspects should be considered when developing a tool. Firstly, data minimization should be prioritized, which involves collecting only essential data to achieve the objectives [45]. By collecting fewer data, the risk to victims and vulnerable groups in society can be minimized, simplifying the process of ensuring data privacy [26].

In addition, stakeholders should refrain from collecting personal information during data collection whenever possible. In cases where collecting personal information is unavoidable, it is crucial to erase or mask such information using anonymization and pseudonymization techniques [55, 76, 107]. These techniques are particularly useful when handling sensitive data related to human trafficking and are commonly employed in fields such as healthcare to protect patient data [5, 58].

Stakeholders should also perform periodic audits of tools and their data management practices to ensure continuous compliance with privacy standards and detect and address vulnerabilities (Author Paper Under Review). These audits provide a means to evaluate whether the tool adequately upholds the privacy standards put in place and offer an opportunity to pinpoint any shortcomings or potential threats connected to the tool's data-handling processes.

6 Challenges and Limitations

At the beginning of our collaboration with Organization A, we hypothesized that they engaged in active collaboration with a diverse range of anti-trafficking stakeholders, particularly law enforcement agencies, due to the scale and severity of the problem. However, through direct observation, we identified a significant lack of collaboration stemming from various challenges discussed in Section 4.2.4. This unexpected discovery prompted us to explore not only the barriers to collaboration, but also the broader challenges faced by Organization A.

A primary challenge during this study was maintaining the integrity of Organization A's work, with concerns about publishing findings at an academic conference. This reflects a larger issue in the field, where there is apprehension about openly sharing investigative methods for fear that traffickers could adapt their strategies.

We obtained permission from Organization A to share findings while anonymizing crucial details such as platform names and the country of focus. This approach aimed to prevent inadvertent identification and alleviate concerns. Balancing the exchange of crucial insights with the academic community and protecting the effectiveness of stakeholders' efforts is a major challenge in this field.

A major limitation of this study was its focus on a single online platform in one African nation. Although traffickers utilize various online channels, Organization A primarily targets Platform A for its work as previously mentioned. Unfortunately, our efforts to partner with other organizations in this field proved challenging. However, it is crucial for the academic community to expand investigations to encompass a wider range of online platforms, such as job boards, and to explore different countries. Research across various platforms and regions will yield a more comprehensive understanding of online human trafficking activities.

7 Conclusion

This study employed in-person and digital observations, semi-structured expert interviews, and case study analyses to investigate human trafficking recruitment on social media platforms. The findings highlight the tactics used by human traffickers on these platforms, and the indicators associated with suspicious job advertisements and profiles. Furthermore, the study discusses the methodologies that anti-trafficking organizations employ to identify and investigate fraudulent job advertisements. It makes four primary contributions: (1) it sheds light on the strategies employed by traffickers, particularly their use of deceptive fraudulent job advertisements as a tool for online recruitment, and examines the nature of these advertisements; (2) it provides an empirical understanding of the techniques used by organizations combating human trafficking to investigate recruitment activities on social media platforms, as well as the obstacles they encounter; (3) it highlights the ethical considerations and difficulties associated with developing software to combat trafficking, offering suggestions to address these concerns; and (4) it outlines design recommendations for potential features that could be implemented on social media platforms to reduce the risk of human trafficking recruitment. The purpose of this study was to contribute to the existing body of knowledge on developing safer digital environments and combating online human trafficking recruitment. Future research should focus on expanding the scope of this investigation to other regions where online recruitment through fraudulent jobs is a significant issue. In addition, other online platforms, including job boards, which are also hubs for fraudulent job listings used in human trafficking, should be examined to further enhance our understanding of this phenomenon.

Acknowledgments

We extend our gratitude to the anti-trafficking organization for granting us the opportunity to witness their operations and speak with their staff during our research.

References

- [1] 5Rights Foundation. 2023. 5Rights African Union approves pioneering Child Online Safety and Empowerment Policy, leveraging 5Rights Toolkit. <https://5rightsfoundation.com/in-action/african-union-adopts-pioneering-child-online-safety-and-empowerment-policy-leveraging-5rights-toolkit.html#:~:text=The%20African%20Union%20Ministerial%20Meeting,empowerment%20by%20design%20and%20default.>
- [2] ActiveFence. [n. d.]. Compliance and Regulation. <https://www.activefence.com/compliance-regulation/>
- [3] ActiveFence. 2023. Advancing Trust Safety. (2023).
- [4] African Union. 2024. Africa Has Become The First Region in The World to Implement a Child Online Safety and Empowerment Policy | African Union. <https://au.int/en/pressreleases/20240523/child-online-safety-and-empowerment-policy-africa-union#:~:text=The%20policy%2C%20and%20its%20proposed,digital%20society%20and%20economy%20with>

- [5] Mishall Al-Zubaidie, Zhongwei Zhang, and Ji Zhang. 2019. PAX: Using pseudonymization and anonymization to protect patients' identities and data in the healthcare system. *International Journal of Environmental Research and Public Health* 16, 9 (5 2019). doi:10.3390/ijerph16091490
- [6] Caitlin Allen. 2019. The Role of the Internet on Sex Trafficking - International Observatory of Human Rights. <https://observatoryihr.org/blog/the-role-of-the-internet-on-sex-trafficking/>
- [7] Issa Alsmadi and Keng Hoon Gan. 2019. Review of short-text classification. 155–182 pages. doi:10.1108/IJWIS-12-2017-0083
- [8] Hamidreza Alviri, Paulo Shakarian, and J. E. Kelly Snyder. 2016. A non-parametric learning approach to identify online human trafficking. *IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI 2016* 2000, Tvp (2016), 133–138. doi:10.1109/ISI.2016.7745456
- [9] Brittany Anthony. 2018. On-Ramps, Intersections, Routes: and Exit A Roadmap for Systems and Industries to Prevent and Disrupt Human Trafficking. (2018).
- [10] Hassan Marzoughi Ardakani. 2020. LSU Digital Commons Identifying Human Trafficking Networks in Louisiana by Using Authorship Attribution and Network Modeling IDENTIFYING HUMAN TRAFFICKING NETWORKS IN LOUISIANA BY USING AUTHORSHIP ATTRIBUTION. *LSU Doctoral Dissertations* 5274 (2020).
- [11] M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. D'Aquin, K. Srinivas, P. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, and S. Staab. 2015. The semantic Web - ISWC 2015: 14th international semantic web conference bethelehem, PA, USA, October 11-15, 2015 proceedings, part II. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9367 (2015), 205–221. doi:10.1007/978-3-319-25010-6
- [12] Denni Arli. 2017. Does Social Media Matter? Investigating the Effect of Social Media Features on Consumer Attitudes. *Journal of Promotion Management* 23, 4 (7 2017), 521–539. doi:10.1080/10496491.2017.1297974
- [13] Renata Arronte. 2018. A partner in TRAFFICKING: THE ROLE OF INTERNET TECHNOLOGIES IN THE FACILITATION OF HUMAN TRAFFICKING. December (2018).
- [14] Coryn Barclay. 2018. *Semi-Structured Interviews*. Technical Report. http://sociology.fas.harvard.edu/files/sociology/files/interview_strategies.pdf
- [15] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. (10 2018). <http://arxiv.org/abs/1810.01943>
- [16] Tolga Bolukbasi, KaiWei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. (2016).
- [17] Hannah E. Britton and Laura A. Dean. 2014. Policy Responses to Human Trafficking in Southern Africa: Domesticating International Norms. *Human Rights Review* 15, 3 (2014), 305–328. doi:10.1007/s12142-014-0303-9
- [18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. (5 2020). <http://arxiv.org/abs/2005.14165>
- [19] Louis Bruijn. 2020. Inter-Annotator Agreement (IAA). Pair-wise Cohen kappa and group Fleiss'... | by Louis de Bruijn | Towards Data Science. <https://towardsdatascience.com/inter-annotator-agreement-2f46c6d37bf3>
- [20] Imogen Calderwood. 2018. Human Traffickers Using Facebook to Lure Migrants Into 'Trips'. <https://www.globalcitizen.org/en/content/human-trafficking-facebook-adverts-refugees/>
- [21] Paolo Campana. 2022. *Online and technology-facilitated trafficking in human beings Summary and recommendations Group of Experts on Action against Trafficking in Human Beings*. Technical Report.
- [22] Youngjin Chae and Thomas Davidson. 2023. LARGE LANGUAGE MODELS FOR TEXT CLASSIFICATION: FROM ZERO-SHOT LEARNING TO FINE-TUNING Large language models for text classification. Technical Report.
- [23] Lura Chamberlain. 2019. *FOSTA: A Hostile Law with a Human Cost*. Technical Report. 2171 pages.
- [24] Man pui Sally Chan, Christopher R. Jones, Kathleen Hall Jamieson, and Dolores Albarracin. 2017. Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science* 28, 11 (11 2017), 1531–1546. doi:10.1177/0956797617714579
- [25] Diane Cottrell. 2023. The Impact of Social Media, the Internet, and Legislation on Online Minor Sex Trafficking. *Journal of Digital Life and Learning* 3, 2 (7 2023), 18–45. doi:10.51357/jdll.v3i2.226
- [26] DataGrail. 2023. How to Implement Data Minimization | DataGrail. <https://www.datagrail.io/blog/data-privacy/how-to-implement-data-minimization/>
- [27] Julia Deeb-Swihart. 2022. DEEB-SWIHART-DISSERTATION-2022. (2022).

- [28] Julia Deeb-Swihart, Alex Endert, and Amy Bruckman. 2022. Ethical Tensions in Applications of AI for Addressing Human Trafficking: A Human Rights Perspective. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (11 2022), 1–29. doi:10.1145/3555186
- [29] Julia Deeb-Swihart, Alex Endert, and Amy Bruckman. 2022. Ethical Tensions in Applications of AI for Addressing Human Trafficking: A Human Rights Perspective. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (11 2022). doi:10.1145/3555186
- [30] Us Department of State. 2023. *OFFICE TO MONITOR AND COMBAT T RAFFICKING IN PERSONS Online Recruitment of Vulnerable Populations for Forced Labor*. Technical Report. <https://www.state>.
- [31] Dhaliwal Jasdev. [n. d.]. Online Job Scams – TikTokers Tell Their Stories, with a Warning | McAfee Blog. <https://www.mcafee.com/blogs/internet-security/online-job-scams-tiktokers-tell-their-stories-with-a-warning/>
- [32] Katherine Dorton. 2019. THE EVOLUTION OF HUMAN TRAFFICKING: THE USE OF SOCIAL MEDIA AND FINANCIAL INSTITUTIONS. December (2019).
- [33] 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 (2015), 65–85. doi:10.1080/23322705.2015.1015342
- [34] Egnite. 2021. Data Auditing – Improve Data Quality | Egnite. <https://www.egnyte.com/guides/governance/data-auditing>
- [35] European Commission. [n. d.]. The Digital Services Act. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en
- [36] Europol. 2014. Intelligence Notification: Trafficking in Human Beings and the Internet. October (2014), 1–3. <https://www.europol.europa.eu/publications-documents/trafficking-in-human-beings-and-internet>
- [37] Camilla Fabbri, Heidi Stöckl, Cathy Zimmerman, Katharine Jones, Harry Cook, Claire Galez-Davis, Naomi Grant, and Yuki Lo. 2023. Labor Recruitment and Human Trafficking: Analysis of a Global Trafficking Survivor Database. doi:10.1177/01979183221139120
- [38] Amy Fedele. 2021. Human Trafficking Trends in Sub-Saharan Africa (Infographic) | ASEC-SLDI News. <http://asec-sldi.org/news/current/human-trafficking-sub-saharan-africa/>
- [39] Kyleigh Feehs and Alyssa Currier Wheeler. 2020. Federal Human Trafficking Report. (2020).
- [40] Mackenzie Flynn. 2021. FOSTA-SESTA and its impact on sex workers - AIDS United. <https://aidsunited.org/fosta-sesta-and-its-impact-on-sex-workers/>
- [41] Freedom Signal. [n. d.]. Freedom Signal. <https://freedomsignal.org/>
- [42] João Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. doi:10.1145/2523813
- [43] Government Gazette and Government Notice. 2022. *Films and Publications Act: Regulations: Amendment*. Technical Report. www.gpwonline.co.za
- [44] Global Slavery Index. 2018. Africa | Region Highlights. <https://www.globallslaveryindex.org/2018/findings/regional-analysis/africa/>
- [45] Abigail Goldstein, Gilad Ezov, Ron Shmelkin, Micha Moffie, and Ariel Farkash. 2022. Data minimization for GDPR compliance in machine learning models. *AI and Ethics* 2, 3 (8 2022), 477–491. doi:10.1007/s43681-021-00095-8
- [46] GOV.UK. 2023. A guide to the Online Safety Bill - GOV.UK. <https://www.gov.uk/guidance/a-guide-to-the-online-safety-bill>
- [47] V Greiman and C Bain. 2013. The Emergence of Cyber Activity. *International Journal of Cyber Warfare and Terrorism (IJCWT)* 12, 2 (2013), 41–49.
- [48] He He, Sheng Zha, and Haohan Wang. 2019. Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual. (8 2019). <http://arxiv.org/abs/1908.10763>
- [49] Myriam Hernandez-Alvarez. 2019. Detection of possible human trafficking in twitter. *Proceedings - 2019 International Conference on Information Systems and Software Technologies, ICI2ST 2019* (2019), 187–191. doi:10.1109/ICI2ST.2019.00034
- [50] Kristin Houser. 2019. Human Traffickers Are Selling Slaves on Instagram | Futurism. <https://futurism.com/human-traffickers-selling-slaves-instagram>
- [51] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. doi:10.1145/3290605.3300637
- [52] Marisa Hultgren, Murray E. Jennex, John Persano, and Cezar Ornatowski. 2016. Using knowledge management to assist in identifying human sex trafficking. *Proceedings of the Annual Hawaii International Conference on System Sciences* 2016-March, March 2019 (2016), 4344–4353. doi:10.1109/HICSS.2016.539
- [53] Michelle Ibanez and Daniel D. Suthers. 2014. Detection of domestic human trafficking indicators and movement trends using content available on open internet sources. *Proceedings of the Annual Hawaii International Conference on*

- System Sciences* (2014), 1556–1565. doi:10.1109/HICSS.2014.200
- [54] Iguazio. [n. d.]. What is Concept Drift . <https://www.iguazio.com/glossary/concept-drift/>
- [55] Information Commissioner's Office. 2022. *Chapter 3: pseudonymisation Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance*. Technical Report.
- [56] Fereshte Khani and Percy Liang. 2020. Removing Spurious Features can Hurt Accuracy and Affect Groups Disproportionately. (12 2020). <http://arxiv.org/abs/2012.04104>
- [57] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan Van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, and Artem Sokolov. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Bonaventure F. P. Dossou* 12 (2022), 47. doi:10.1162/tacl
- [58] Mostafa Langarizadeh, Azam Orooji, and Abbas Sheikhtaheri. 2018. Effectiveness of anonymization methods in preserving patients' privacy: A systematic literature review. In *Studies in Health Technology and Informatics*, Vol. 248. IOS Press BV, 80–87. doi:10.3233/978-1-61499-858-7-80
- [59] Mark Latonero. 2012. The Rise of Mobile and the Diffusion of Technology-Facilitated Trafficking. *SSRN Electronic Journal* November (2012). doi:10.2139/ssrn.2177556
- [60] Mark Latonero and Lauren Movius. 2011. Center on Communication Leadership & Policy Human Trafficking Online The Role of Social Networking Sites and Online Classifieds. *Research Assistant Tala Mohebi, M.A., Research Associate* September (2011). http://technologyandtrafficking.usc.edu/files/2011/09/HumanTrafficking_FINAL.pdf
- [61] Mary Graw Leary. 2014. Fighting Fire with Fire : Technology in Child Sex Trafficking " From low-tech methods such as prostituting minors at truck stops , to high-tech or trade ." 21 (2014).
- [62] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (7 2019). <http://arxiv.org/abs/1907.11692>
- [63] Rachel Marshall. 2016. *Sex Workers and Human Rights, A Critical Analysis of Laws Sex Work*. Technical Report. 2016–2017 pages. <http://perma.cc/A8BDZNUR>].
- [64] Cameron Martel and David G. Rand. 2023. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. doi:10.1016/j.copsyc.2023.101710
- [65] Michael McKenna. [n. d.]. Bias in AI: How to Mitigate Bias in AI Systems | Toptal®. <https://www.toptal.com/artificial-intelligence/mitigating-ai-bias>
- [66] Meta. [n. d.]. Community Standards Enforcement | Transparency Center. <https://transparency.meta.com/reports/community-standards-enforcement/fake-accounts/facebook/#content-actioned>
- [67] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. 2020. Moments in Time Dataset: One Million Videos for Event Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 2 (2 2020), 502–508. doi:10.1109/TPAMI.2019.2901464
- [68] Dhiraj Murthy. 2008. Digital ethnography: An examination of the use of new technologies for social research. 837–855 pages. doi:10.1177/0038038508094565
- [69] National Computer and Cybercrime Coordination Committee. [n. d.]. The Computer Misuse and Cybercrimes Act 2018 | NC4. <https://nc4.go.ke/the-computer-misuse-and-cybercrimes-act-2018/#:~:text=CMCA%20provides%20for%20offences%20relating,matters%2C%20and%20for%20connected%20purposes>.
- [70] NETalent. 2024. NETalent African Union approves a pioneering Child Online Safety and Empowerment Policy, leveraging 5Rights Toolkit. <https://www.netalent.pro/views/EN/article.html?id=240122>
- [71] Newsome Melton. 2020. What Is the Process of Sex and Human Trafficking? <https://www.newsomelaw.com/sex-trafficking-lawyer/what-is-the-process-of-sex-and-human-trafficking/>
- [72] OneTrust Data Guidance. 2022. Australia: The scope of the Online Safety Act | Insights | DataGuidance. <https://www.dataguidance.com/opinion/australia-scope-online-safety-act>
- [73] Browne Onuoha. 2011. The state human trafficking and human rights issues in Africa. *Contemporary Justice Review: Issues in Criminal, Social, and Restorative Justice* 14, 2 (6 2011), 149–166. doi:10.1080/10282580.2011.565973
- [74] OSCE. 2020. *Leveraging innovation to fight trafficking in human beings : A comprehensive analysis of technology tools*.
- [75] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. doi:10.1016/j.patter.2021.100336
- [76] Ambika Pawar, Swati Ahirrao, Mr Prathamesh, and P Churi. [n. d.]. *Anonymization Techniques for Protecting Privacy: A Survey*. Technical Report.
- [77] Chris (Behavioural researcher) Perry, Krishna Chhatralia, Dom Damesick, Sylvie Hobden, Leanora Volpe, and Health Foundation (Great Britain). 2015. *Behavioural insights in health care : nudging to reduce inefficiency and waste*. 73 pages.
- [78] Eliska Pirkova. 2022. The Digital Services Act: your guide to the EU's new content moderation rules - Access Now. <https://www.accessnow.org/digital-services-act-eu-content-moderation-rules-guide/>

- [79] Progetto Tenda, Danish Refugee Council, Nesta Italia, LABC Srl, and CWEP. 2021. Free2Link Report. (2021).
- [80] Prostasia Foundation. [n. d.]. Support the SAFE SEX Workers Study Act - Prostasia Foundation. <https://prostasia.org/campaign/support-the-safe-sex-workers-study-act/>
- [81] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. [n. d.]. *Language Models are Unsupervised Multitask Learners*. Technical Report. https://github.com/openai/gpt-2/blob/master/language_model/unsupervised_multitask_learners.pdf
- [82] Rephrain. 2024. REPHRAIN – National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online. <https://www.rephrain.ac.uk/>
- [83] Ines Roldós. [n. d.]. Major Challenges of Natural Language Processing (NLP). <https://monkeylearn.com/blog/natural-language-processing-challenges/>
- [84] Aja Romano. 2018. FOSTA-SESTA, a law intended to curb sex trafficking, threatens the internet’s future - Vox. <https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom>
- [85] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale. Technical Report. <http://winogrande.allenai.org>
- [86] Siddhartha Sarkar. 2015. Use of technology in human trafficking networks and sexual exploitation: A cross-sectional multi-country study. *Transnational Social Review* 5, 1 (2015), 55–68. doi:10.1080/21931674.2014.991184
- [87] Seldon. 2021. Machine Learning Concept Drift - What is it and Five Steps to Deal With it - Seldon. <https://www.seldon.io/machine-learning-concept-drift>
- [88] Abby Seneor and Matteo Mezzanotte. 2022. Open source data science: How to reduce bias in AI | World Economic Forum. <https://www.weforum.org/agenda/2022/10/open-source-data-science-bias-more-ethical-ai-technology/>
- [89] Shabnoor Siddiqui. 2016. *Social Media its Impact with Positive and Negative Aspects*. Technical Report 2. 71–75 pages. www.ijcat.com
- [90] Daniel Ribeiro Silva, Andrew Philpot, Abhishek Sundararajan, Nicole Marie Bryan, and Eduard Hovy. 2014. Data integration from open internet sources and network detection to combat underage sex trafficking. *ACM International Conference Proceeding Series* (2014), 86–90. doi:10.1145/2612733.2612746
- [91] City Simon. 2018. Perma | On Backpage | Tits and Sass. <https://perma.cc/39AT-6MKT>
- [92] Ellie Louise Simonson. 2021. Semi-Supervised Classification of Social Media Posts : Identifying Sex-Industry Posts to Enable Better Support for Those Experiencing Sex-Trafficking by Semi-Supervised Classification of Social Media Posts : Identifying Sex-Industry Posts to Enable Better. (2021).
- [93] Devin Soni. 2022. Feedback Loops in Machine Learning Systems | by Devin Soni | Towards Data Science. <https://towardsdatascience.com/feedback-loops-in-machine-learning-systems-701296c91787>
- [94] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. (9 2020). <http://arxiv.org/abs/2009.10795>
- [95] Athanassia Sykiotou. 2007. Trafficking in human beings: Internet recruitment. *Europe* (2007).
- [96] Rende Lisa Taylor and Mark Latonero. 2018. Updated Guide to Ethics and Human rights in Anti-Human Trafficking: Ethical Standards for Working with Migrant Workers and Trafficked Persons in the Digital Age. (2018). <http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf>
- [97] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2020. Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision. (4 2020). <http://arxiv.org/abs/2004.09034>
- [98] The UN Agency for Digital Technologies. 2021. Are African countries doing enough to ensure cybersecurity and Internet safety? - ITU. <https://www.itu.int/hub/2021/09/are-african-countries-doing-enough-to-ensure-cybersecurity-and-internet-safety/>
- [99] Thorn. 2018. Survivor Insights:The Role of Technology in Domestic Minor Sex Trafficking JANUARY. Thorn (2018).
- [100] Edmund Tong, Cara Jones, Amir Zadeh, and Louis Philippe Morency. 2017. Combating human trafficking with deep multimodal models. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 1 (2017), 1547–1556. doi:10.18653/v1/P17-1142
- [101] UNICEF. 2003. *Trafficking in human beings, especially women and children, in Africa*. UNICEF Innocenti Research Center. 72 pages.
- [102] United Nations. 2000. Protocol to Prevent, Suppress and Punish Trafficking in Persons Especially Women and Children, supplementing the United Nations Convention against Transnational Organized Crime | OHCHR. <https://www.ohchr.org/en/instruments-mechanisms/instruments/protocol-prevent-suppress-and-punish-trafficking-persons>
- [103] United Nations Office on Drugs and Crime. 2018. *Global Report on Trafficking in Persons*. Technical Report.
- [104] United Nations Office on Drugs and Crime. 2020. *Global Report on Trafficking in Persons*. (2020).
- [105] US Department of State. 2019. *TRAFFICKING IN PERSONS REPORT*. (2019).
- [106] US Department of State. 2023. *Report to Congress on 2023 Trafficking in Persons Interim Assessment Pursuant to the Trafficking Victims Protection Act - United States Department of State*. <https://www.state.gov/report-to-congress-on-2023-trafficking-in-persons-interim-assessment-pursuant-to-the-trafficking-victims-protection-act/>

- act/#::~text=The%20Trafficking%20Victims%20Protection%20Act%20of%202000%20as%20amended%20(TVPA,the%20countries%20on%20that%20list.
- [107] Liudmyla Vasylieva. 2021. *Anonymisation and pseudonymisation of textual documents*. Technical Report.
 - [108] Ada Volodko, Ella Cockbain, and Bennett Kleinberg. 2020. “Spotting the signs” of trafficking recruitment online : exploring the characteristics of advertisements targeted at migrant job-seekers. (2020).
 - [109] S Walby, B Apitzsch, J. E. Armstrong, S Balderston, K Szmagalska-Follis, Kelly L Francis, B. J., C. A May-Chahal, A Rashid, K. Shire, J. Towers, and M Tunte. 2016. Study on the gender dimension of trafficking in human beings. (2016), 18–31.
 - [110] Angelina Wang. 2020. *REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets*. Technical Report. <https://github.com/princetonvisualai/revise-tool>.
 - [111] Hao Wang, Congxing Cai, Andrew Philpot, Mark Latonero, Eduard H. Hovy, and Donald Metzler. 2012. Data integration from open internet sources to combat sex trafficking of minors. *ACM International Conference Proceeding Series* (2012), 246–252. doi:10.1145/2307729.2307769
 - [112] Longshaokan Wang, Eric Laber, Yeng Saanchi, and Sherrie Caltagirone. 2020. Sex trafficking detection with ordinal regression neural networks. *arXiv* April 2018 (2020).
 - [113] Peter Waters. 2023. The perils of feedback loops in machine learning: predictive policing | Gilbert + Tobin Lawyers: Law Firm in Sydney, Melbourne & Perth. <https://www.gtlaw.com.au/knowledge/perils-feedback-loops-machine-learning-predictive-policing>
 - [114] Jessica Whitney. 2017. KM VS HUMAN TRAFFICKING : AN EXPLORATORY STUDY ON USING EMOJIS FOR A KNOWLEDGE DRIVEN APPROACH TO IDENTIFYING ONLINE HUMAN SEX TRAFFICKING Presented to the Faculty of San Diego State University In Partial Fulfillment of the Requirements for the Degree Mas. (2017).
 - [115] Jessica Whitney, Murray Jennex, Aaron Elkins, and Eric Frost. 2018. Don’t Want to Get Caught? Don’t Say It: The Use of EMOJIS in Online Human Sex Trafficking Ads. *Proceedings of the 51st Hawaii International Conference on System Sciences* (2018), 4273–4282. doi:10.24251/hicss.2018.537
 - [116] Maarit2 Widmann. 2020. Cohen’s Kappa: what it is, when to use it, how to avoid pitfalls | KNIME. <https://www.knime.com/blog/cohens-kappa-an-overview>
 - [117] Kate Williams. 2022. Participant Observation 101: Definition, Types, Uses, Examples. <https://surveysparrow.com/blog/participant-observation/#:::text=Passive%20participant%20observation,done%20without%20the%20participants%20knowledge>.
 - [118] Cao Xiao, David Mandell Freeman, and Theodore Hwa. 2015. Detecting clusters of fake accounts in online social networks. *AISeC 2015 - Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, co-located with CCS 2015* (2015), 91–102. doi:10.1145/2808769.2808779
 - [119] Siyi Zhou, Jiankun Peng, and Emilio Ferrara. 2024. *Tracing the Unseen: Uncovering Human Trafficking Patterns in Job Listings*. Technical Report. <https://www.chineseinla.com/user/id>
 - [120] Jessica Zhu, Lin Li, and Cara Jones. 2019. Identification and detection of human trafficking using language models. *Proceedings of the 2019 European Intelligence and Security Informatics Conference, EISIC 2019* (2019), 24–31. doi:10.1109/EISIC49498.2019.9108860

Received January 2024; revised July 2024; accepted October 2024