# Investigating and Detecting Human Trafficking Recruitment on Online Platforms Using a Responsible Innovation Approach




Towera Moyo

St John's College

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Trinity 2024

This thesis is dedicated to
my lovely siblings, Khumbo and Kutemwa, and
and to my parents, who have passed away.
Most importantly, to God, without whom I could not have made it this far.

# Acknowledgements

I thank God for being with me throughout this journey, for giving me the strength to overcome challenges, and for guiding me to achieve this significant milestone.

I extend my heartfelt thanks to my supervisor, Professor Marina Jirotka, for her unwavering support throughout my thesis and belief in my research topic. Her guidance was crucial to the success of my DPhil, and her encouragement has been equally impactful in my career endeavours. I also thank my other supervisor, Dr. Omer Gunes, for joining midway through my DPhil and offering his guidance and resources to see it through to a successful completion. My appreciation also goes to Dr. Helena Webb, who was my supervisor during the first year but continued to offer support for my DPhil in various ways, including career prospects, even after her departure.

I also extend my appreciation to the various participants and individuals who have shaped my work. A special thank you to Organisation A for providing me with data, granting me the opportunity to observe their work, and allowing me to interview their staff. Special thanks to the former head of data science for approving my request for collaboration and providing access to the data used for this thesis, and to the director for their continued support and approval to engage with both data and staff. I am particularly grateful to the project manager and the investigators, whose invaluable assistance in responding to my many questions, participating in interviews, and attending workshops was instrumental in my work.

Special thanks go to my family, particularly my father, **Mr Humphrey Bright Moyo**, whose lifelong support, love, and assistance in securing funding for this PhD were vital to my success. I am also very grateful to my brother, Khumbo Moyo, for his constant encouragement and support throughout the DPhil. My heartfelt appreciation extends to the Moyo and Sibande families, including the Mia Familia cousins, for their love and support, with a special mention of my cousin, Chawezi Sibande, who helped me review my DPhil proposal and has, on many occasions, assisted in preparing for important DPhil presentations.

I am incredibly thankful to my friends, especially Muhammad, Brian, Nadja, Hye Min, Sandra, Manal, Seluleko, Cephas, Nerissa, and Marah, for being amazing support systems and for making this experience so much better. I am also very grateful to Dr. Frank Mtumbuka, who provided assistance in various capacities,

including proofreading. Heartfelt thanks go to my dear friend, Dr. Andile Dube, who supported me throughout my DPhil, from proofreading my work to helping with viva presentations and job applications despite his busy schedule. A major thank you goes to Dr. Albert Sharra, who proofread the entire thesis and was always available to assist.

I would also like to acknowledge my colleagues from 39a, who shared many laughs with me and offered support in various ways over the past four years. Special thanks to Ross Gales, Keri Grieman, and Josimar Mendes for reviewing parts of my work on various occasions. I also extend my sincere thanks to Carolyn Ten Holter and Daniel Omeiza, who were consistently helpful and supportive, always willing to proofread my work, including academic papers, and answer any questions I had.

Finally, I am profoundly grateful to the Rhodes Trust for funding this DPhil thesis. This achievement would not have been possible without their support. The scholarship not only provided financial assistance, but also supported me through difficult times. Special thanks go to the registrar, Mary Eaton, for always lending a listening ear and helping to resolve various issues related to my thesis.

# Abstract

Human trafficking affects more than 40 million people in the world. In recent years, human traffickers have shifted from conventional physical methods and are increasingly using online platforms to identify and exploit victims for various forms of abuse. These harms can affect victims physically, emotionally, and psychologically. This thesis examines human trafficking recruitment on online platforms and explores the application of artificial intelligence to detect fraudulent job advertisements used for victim recruitment, while considering guidelines that mitigate potential unintended negative consequences. This research was carried out in four phases. The first phase was a qualitative study that explored the implications of developing and using anti-trafficking tools. Guided by the principles of responsible research and innovation, this phase identified challenges and the intended and unintended consequences of deploying anti-trafficking technology. It provided recommendations for addressing these challenges and proposed a framework to mitigate potential negative consequences, informing the subsequent phases of this research. The second phase focused on understanding human trafficking recruitment. Through the analysis and visualisation of two real-world datasets, this phase uncovered patterns, profiles of victims and perpetrators, recruitment indicators, processes, and recurring trends. The third phase was an ethnographic study that investigated how traffickers use social media to recruit victims using fraudulent job advertisements. It also provided insights into the indicators of such advertisements and highlighted the strategies anti-trafficking organisations employ to identify them. In the final phase, a labelled dataset consisting of job advertisements related to human trafficking was developed. This phase included benchmarking traditional machine learning algorithms, pre-trained neural language models, and large language models to classify fraudulent job advertisements. The research found that pre-trained neural language models were the most effective in identifying advertisements related to human trafficking, and these models were subsequently used to label 10,000 unlabelled data points. The research methods employed in these four phases were guided by principles of ethical responsibility, recent advancements in computer science, and a mixed-methods approach. These guiding principles ensured that the selected methods were both scientifically rigorous and socially responsible in addressing the goals of the research. The four phases collectively contributed to a deeper understanding of human trafficking recruitment and demonstrated how technology can be leveraged to help combat this global issue.

# Contents

**List of Figures** **xii**

**List of Tables** **xiv**

**List of Abbreviations** **xv**

**1 Introduction** **1**
- 1.1 The Research . . . 4
- 1.2 Research Question and Objectives . . . 5
  - 1.2.1 Research Questions . . . 5
  - 1.2.2 Research Objectives . . . 6
- 1.3 Methodology . . . 7
- 1.4 Contributions . . . 9
- 1.5 Terminology . . . 11
- 1.6 Publications . . . 12
- 1.7 Thesis Outline . . . 12

**2 Background and Literature Review** **15**
- 2.1 Introduction . . . 16
- 2.2 Overview of Human Trafficking . . . 17
  - 2.2.1 Types of Human Trafficking . . . 17
  - 2.2.2 Human Trafficking Actors . . . 19
  - 2.2.3 Stages of Human Trafficking . . . 22
- 2.3 Overview of Technology-Facilitated Trafficking . . . 24
  - 2.3.1 Online Recruitment . . . 26
  - 2.3.2 Online Exploitation . . . 31
- 2.4 Anti-Trafficking Tools . . . 34
- 2.5 The Impact of Anti-Trafficking Tools . . . 37
- 2.6 Related Work on Technological Solutions to Detect Online Trafficking 39
  - 2.6.1 Data Sources and Acquisition . . . 40
  - 2.6.2 Detecting Interned-based Trafficking using Machine Learning and Natural Language Processing . . . 42
- 2.7 Anti-trafficking Policies and Online Legislation . . . 49

2.8 Summary and Current Limitations . . . . . 51
2.8.1 Limited Studies that Assess the Impact of Anti-Trafficking Tools . . . . . 51
2.8.2 Limited Work on Online Recruitment . . . . . 52
2.8.3 Lack of Anti-trafficking Technologies that Identify Online Recruitment . . . . . 53
2.8.4 Limited Use of State-of-the-art Methods in Natural Language Processing . . . . . 54
2.9 Direction of Thesis . . . . . 55

**3 Assessing Anti-Trafficking Software 57**
3.1 Introduction . . . . . 58
3.2 Methodology . . . . . 59
3.2.1 Responsible Research and Innovation . . . . . 60
3.2.2 Anticipate: Analysing the Consequences of Anti-trafficking Tools . . . . . 61
3.2.3 Reflection: Examination of Motivations and Benefits . . . . 62
3.2.4 Engagement: Involving Diverse Stakeholders . . . . . 63
3.2.5 Act: Integrating Stakeholder Perspectives . . . . . 67
3.2.6 Comparative Approach: US and UK Contexts . . . . . 67
3.3 Research Findings . . . . . 67
3.3.1 Challenges Associated with Developing Anti-Trafficking Tools 68
3.3.2 Positive Consequences of Anti-Trafficking Tools . . . . . 71
3.3.3 Negative Consequences of Anti-trafficking Systems . . . . . 74
3.3.4 Recommendations for Anti-trafficking Tool Users . . . . . 83
3.4 Proposed Framework for Developing Anti-Trafficking Tools: P3D Framework . . . . . 84
3.4.1 Project Planning . . . . . 85
3.4.2 Data Preparation . . . . . 89
3.4.3 Design Phase . . . . . 93
3.4.4 Deployment Phase . . . . . 98
3.4.5 Summary . . . . . 100
3.5 Discussion and Recommendations . . . . . 101
3.5.1 Summary of Findings . . . . . 101
3.5.2 Responsible Research and Innovation Implications . . . . . 102
3.5.3 Recommendations for Policymakers . . . . . 103
3.6 Conclusion . . . . . 104
3.7 Direction of the thesis . . . . . 104

**4 Analysis of Human Trafficking Recruitment in Africa 107**
4.1 Introduction . . . 108
4.2 Background . . . 109
4.2.1 Human Trafficking in Africa . . . 109
4.2.2 Human Trafficking Recruitment . . . 110
4.2.3 Factors that Contribute to Human Trafficking Recruitment in Africa . . . 111
4.3 Organisation A . . . 113
4.3.1 Background . . . 113
4.4 Methodology . . . 118
4.4.1 Data Collection and Preparation . . . 118
4.4.2 Data Visualisation . . . 120
4.5 Results from Analysis of Dataset: Form 1 . . . 121
4.5.1 Overview Analysis: Form 1 . . . 121
4.5.2 General Hypothesis Testing . . . 128
4.6 Results from Analysis of Dataset: Form 2 . . . 132
4.6.1 Overview Analysis: Form 2 . . . 132
4.6.2 Analysis by Country: Form 2 . . . 135
4.6.3 General Hypothesis Testing . . . 147
4.7 Discussion . . . 149
4.7.1 Dataset 1 . . . 149
4.7.2 Dataset 2 . . . 152
4.7.3 Design Implications . . . 153
4.7.4 Direction of Thesis . . . 156
4.8 Conclusion . . . 157

**5 Investigating Fraudulent Job Advertisements for Human Trafficking Online 159**
5.1 Introduction . . . 160
5.2 Methodology . . . 161
5.2.1 Participant and Platform Background . . . 162
5.2.2 Online Observation: Social Media Groups . . . 164
5.2.3 In-person Observation . . . 165
5.2.4 Semi-structured Interviews . . . 166
5.2.5 Case Notes . . . 167
5.2.6 Analysis . . . 167
5.2.7 Ethical Considerations of Digital Observations . . . 168
5.3 Findings . . . 169
5.3.1 Methods Traffickers Employ to Recruit Victims on Social Media 169

5.3.2 Organisation Work: Raising Awareness . . . . . 170
5.3.3 Organisation Work: Tackling Online Human Trafficking Recruitment . . . . . 172
5.3.4 Identifying Suspicious Job Advertisements . . . . . 177
5.3.5 Characteristics of Suspicious profiles . . . . . 184
5.4 Discussion . . . . . 184
5.4.1 Summary . . . . . 184
5.4.2 Analysis of Organisational Work . . . . . 185
5.4.3 Opportunities for the Use of Technology to Identify and Investigate Suspicious Job Posts . . . . . 188
5.4.4 Design Implications for Social Media Platforms . . . . . 192
5.5 Reflection on Fieldwork . . . . . 194
5.6 Conclusion . . . . . 195

**6 Benchmarking AI Techniques to Detect Fraudulent Online Job Ads 196**
6.1 Introduction . . . . . 197
6.2 Human Trafficking Recruitment Dataset . . . . . 199
6.2.1 Data Collection and Preprocessing . . . . . 199
6.2.2 Data Processing . . . . . 200
6.2.3 Data Annotation and Coding Scheme . . . . . 201
6.2.4 1K Dataset Statistics . . . . . 206
6.3 Classification: Benchmarking Various Approaches . . . . . 207
6.3.1 Traditional Machine Learning Approach . . . . . 207
6.3.2 Pre-trained Neural Language Models . . . . . 215
6.3.3 Large Language Models Approach . . . . . 222
6.4 Expert Feedback . . . . . 224
6.5 Trafficking Recruitment 10K Dataset . . . . . 226
6.6 Discussion . . . . . 227
6.6.1 Summary of Findings . . . . . 228
6.6.2 Analysis of Classification Results . . . . . 228
6.6.3 Analysis of Job Advertisements . . . . . 231
6.6.4 Dataset Design Implications . . . . . 233
6.6.5 Future Technical Implications . . . . . 235
6.6.6 Recommendations for Policymakers . . . . . 237
6.7 Conclusion . . . . . 238

**7 Discussion** **240**
7.1 Summary of Findings . . . 241
7.1.1 Phase 1: Assessing Anti-trafficking Software . . . 243
7.1.2 Phase 2: Analysis of Human Trafficking Recruitment in Africa 247
7.1.3 Phase 3: Investigating Fraudulent Jobs for Human Trafficking on Social Media . . . 250
7.1.4 Phase 4: A Machine Learning Approach to Detecting Human Trafficking Recruitment . . . 254
7.1.5 Reflection on Research Questions . . . 256
7.2 Use of P3D Framework . . . 257
7.2.1 Project Planning . . . 257
7.2.2 Data Planning . . . 258
7.2.3 Design and Deployment . . . 259
7.2.4 Summary . . . 259
7.3 Contributions to Work . . . 260
7.3.1 Empirical Insights . . . 260
7.3.2 Design Recommendations and Framework Contributions . . 261
7.3.3 Technological and Methodological Contributions . . . 262
7.4 Limitations and Challenges . . . 263
7.4.1 Phase 1: Assessing Anti-trafficking Software . . . 263
7.4.2 Phase 2: Analysis of Human Trafficking Recruitment in Africa 264
7.4.3 Phase 3: Investigating Fraudulent Jobs for Human Trafficking on Social Media . . . 264
7.4.4 Phase 4: A Machine Learning Approach to Detecting Human Trafficking Recruitment . . . 265
7.5 Challenges of using Computational Methods to Identify and Investigate Recruitment . . . 266
7.5.1 Achieving Accuracy . . . 266
7.5.2 Meeting Regulatory Requirements . . . 267
7.5.3 Application Integration . . . 268
7.5.4 Long Term Performance . . . 268
7.5.5 Feasibility of Developing and Using Anti-trafficking Software 270
7.6 Recommendations for Future Work . . . 270
7.6.1 Broadening the Scope of Research . . . 271
7.6.2 Creating Datasets . . . 272
7.6.3 Comparing Performance with Social Media Data . . . 272
7.6.4 Developing and Deploying Real-World Systems . . . 272
7.6.5 Developing Optimisation Tools for Anti-trafficking Stakeholders in Africa . . . 273

7.6.6 Long-term performance of real-world anti-trafficking tools . . 274
7.6.7 Studies on Increasing User Trust in Anti-trafficking Software 274
7.6.8 Technical Implications of Anti-trafficking Software . . . . . 274
7.6.9 Generalising Findings in other African Countries . . . . . 275
7.6.10 Refining the P3D Framework . . . . . 275
7.7 Reflection and Recommendations . . . . . 276
7.7.1 Academic Community . . . . . 276
7.7.2 Policymakers . . . . . 279
7.7.3 Anti-trafficking Stakeholders . . . . . 280
7.7.4 Final Thoughts . . . . . 281

**A Chapter 3 Supplementary Materials 283**
A.1 Interview Questions . . . . . 283
A.1.1 Developers . . . . . 283
A.1.2 Law Enforcement . . . . . 284
A.1.3 Anti-trafficking Experts . . . . . 285
A.2 Consent Form . . . . . 286
A.3 Images - Codes and Themes . . . . . 288
A.4 Ethics From . . . . . 291

**B Chapter 4 Supplementary Materials 306**
B.1 Data Cleaning Process . . . . . 306
B.1.1 Dataset 1 . . . . . 306
B.1.2 Dataset 2 . . . . . 307

**C Chapter 6 Supplementary Materials 309**
C.1 Annotation Guide . . . . . 309
C.2 Chapter 6: Initial Data Collection Process . . . . . 312
C.2.1 Query Generator . . . . . 312
C.2.2 Data Collection . . . . . 312
C.2.3 Data Preprocessing . . . . . 313

**References 314**

# List of Figures


1.1 Summary of Thesis . . . . . 13

2.1 The Stages of Human Trafficking . . . . . 23

3.1 Chapter 3 Methodology Process . . . . . 62
3.2 P3D Framework . . . . . 85

4.1 Organisation A Process at Border Control Points . . . . . 114
4.2 Total Victims Over the Years . . . . . 122
4.3 Total Victims- Top 10 Nationalities . . . . . 123
4.4 Methods of Intercepting Victims . . . . . 124
4.5 Signs of Vulnerability . . . . . 125
4.6 Methods of Evading Interceptions . . . . . 126
4.7 Control Means - Reasons for Leaving . . . . . 126
4.8 Control Means - Marriage Details . . . . . 127
4.9 Control Means - Job Details . . . . . 127
4.10 Control Means - Conditions of Victims . . . . . 128
4.11 Summary of Cases . . . . . 128
4.12 Total Number of Victims Based on their Nationality . . . . . 133
4.13 The General Recruitment Methods . . . . . 133
4.14 Methods of Enticing Victims into Trafficking Life . . . . . 134
4.15 Uganda Education Statistics . . . . . 135
4.16 Uganda Top Occupations . . . . . 136
4.17 Uganda Main Recruitment Methods . . . . . 136
4.18 Uganda Top Reasons for Leaving . . . . . 137
4.19 Uganda Top Forms of Exploitation . . . . . 137
4.20 Malawi Education Statistics . . . . . 138
4.21 Malawi Top Occupations . . . . . 139
4.22 Malawi Main Recruitment Methods . . . . . 139
4.23 Malawi Top Reasons for Leaving . . . . . 140
4.24 Malawi Top Forms of Exploitation . . . . . 140
4.25 South Africa Education Statistics . . . . . 141
4.26 South Africa Occupations . . . . . 141

4.27 South Africa Main Recruitment Methods . . . . . 142
4.28 South Africa Top Reasons for Leaving . . . . . 142
4.29 South Africa Top Exploitations . . . . . 143
4.30 Kenya Education Statistics . . . . . 144
4.31 Kenya Occupations . . . . . 144
4.32 Kenya Recruitment Methods . . . . . 145
4.33 Kenya Purpose for Leaving . . . . . 145
4.34 Angola Education Statistics . . . . . 146
4.35 Angola Top Occupations . . . . . 146
4.36 Angola Top Reasons for Leaving . . . . . 147

5.1 Chapter 5 Methodology Process . . . . . 162
5.2 Overview of Organisation B's Workflow . . . . . 172
5.3 Indicator Showing Mass Recruitment . . . . . 179
5.4 Example of a Job Advertisement - Two Posts Advertising Multiple Roles . . . . . 179
5.5 Indicator Showing Writing Style . . . . . 180
5.6 Example of Job Advertisements - Indicator Offering Attractive Packages 182

6.1 Top 20 Unigrams from Job Advertisements Labelled as "High Risk of Human Trafficking" and "Suspicious" (1K Dataset) . . . . . 206
6.2 Overview of Traditional Machine Learning Approach . . . . . 207
6.3 Overview of Pre-trained Neural Language Models Approach . . . . 216
6.4 Overview of Large Language Models Approach . . . . . 222
6.5 Top 20 Unigrams from Job Advertisements Labelled as 'High Risk of Human Trafficking' and 'Suspicious' (10K dataset) . . . . . 227

7.1 Summary of Thesis . . . . . 244

A.1 First Level Coding . . . . . 288
A.2 First Level Coding . . . . . 289
A.3 Chapter 4 - Final Level Coding . . . . . 290
A.4 Study 6 - Coding for Observations and Interviews . . . . . 291

# List of Tables


2.1 A Summary of Anti-trafficking Software . . . . . 35

3.1 Demographic Information of Participants . . . . . 65
3.2 A Summary of Themes from the Interviews . . . . . 68

5.1 Themes and Indicators . . . . . 178

6.1 Kappa Scores (Fendler et al., 2017) . . . . . 203
6.2 Label Statistics . . . . . 206
6.3 Word Embeddings . . . . . 215
6.4 Performance Comparison of Different Algorithms across Multiple Experiments. . . . . 215
6.5 Performance Comparison of BERT, AlBERT, and RoBERTa on the 1K Dataset. . . . . 221
6.6 Performance Comparison of GPT-3.5 and GPT-4 on the 1K Dataset. 224
6.7 Validating Roberta Results . . . . . 226
6.8 Label Statistics . . . . . 227

# List of Abbreviations

**US** . . . . . United States of America.

**UK** . . . . . United Kingdom.

**EU** . . . . . European Union.

**OSCE** . . . . . Organisation for Security and Co-operation in Europe

**FOSTA** . . . . Fight Online Sex Trafficking Act

**SESTA** . . . . Stop Enabling Sex Traffickers Act

**LLMs** . . . . . Large Language Models

**PLMs** . . . . . Pre-trained Neural Language Models

**NLP** . . . . . Natural Language Processing

**ML** . . . . . Machine Learning

**BPE** . . . . . Byte-Pair Encoding

**ICT** . . . . . Information and Communication Technology

**RRI** . . . . . Responsible Research and Innovation

# 1 Introduction

Human trafficking, often referred to as modern slavery, is a widespread problem that affects millions of individuals, subjecting them to severe forms of exploitation, including forced labour, sexual exploitation, and debt bondage. It is also one of the fastest growing criminal industries globally (International Labour Organisation et al., 2022).

Although human trafficking affects people of all genders, women and girls constitute the majority of victims, with reports indicating that they are subjected to violence at a rate that is three times higher than that experienced by men (United Nations Office on Drugs and Crime, 2023). Women and girls are mainly exploited for commercial sex work. The sex trafficking industry is estimated to generate a profit of approximately $99 billion (Miller, 2014; Dorton, 2019; OSCE, 2020; Satyarthi, 2019), making it one of the largest criminal enterprises in the world.

Although human trafficking is primarily a real-world crime, technological advancements have enabled traffickers to adopt new methods to identify and exploit vulnerable individuals. Online platforms, especially social media, adult websites, and classified ads, have become prevalent tools for traffickers to recruit and exploit victims. The Internet allows traffickers to expand their operations beyond geographical boundaries, reaching individuals who would otherwise have been inaccessible (Dorton, 2019; Latonero and Movius, 2011). Human trafficking is thriving on online platforms due to the increasing global use of the Internet,

widespread adoption of technology in developing countries, and the increased dependence on technology exacerbated by the pandemic (Department of State, 2023; Progetto Tenda et al., 2021).

The Internet is used in the initial stages of human trafficking, specifically in the recruitment phase. Two primary methods are commonly employed for online recruitment: posting fraudulent job advertisements and establishing false relationships through grooming (Anthony, 2018). In the former, traffickers deceive unsuspecting individuals with job offers, whereas in the latter, they establish romantic or personal relationships through grooming. Fraudulent job advertisements have been shown to be the most prevalent form of online recruitment in several countries. More than 50% of the human trafficking cases prosecuted in countries such as the United States (US) included fraudulent job offers to recruit people (Feehs and Wheeler, 2020; Ramchandani et al., 2021; United Nations Office on Drugs and Crime, 2020). Human traffickers predominantly use social media platforms because of their ability to operate anonymously and gain easy access to millions of vulnerable individuals seeking better economic opportunities. In 2018, UK officials discovered 539 pages on social media platforms, primarily Facebook, which allowed individuals to travel to Europe (Calderwood, 2018).

Once victims are recruited, the Internet is used for their exploitation. Various organisations have identified websites employed by traffickers for exploitation. Shared Hope International (SHI) identified more than 5,000 websites that advertise victims, while Sykiotou identified an even higher number of 128,000 suspected websites (Latonero and Movius, 2011; Sykiotou, 2007). These examples illustrate the significant role technology plays in the facilitation of human trafficking. The form of human trafficking that is most significantly affected by online technologies is sex trafficking. Human traffickers use online platforms to connect victims with buyers through webcams, disseminate pornographic content, and advertise victims for sexual services (OSCE, 2020; Latonero and Movius, 2011; United Nations Office on Drugs and Crime, 2018; Europol, 2014; Lockett, 2018). Numerous victims, including millions of children, are currently subjected to these online exploitations

(Mitchell and Boyd, 2014). These exploitations can cause physical, emotional, and psychological harm, highlighting the need for preventive measures to protect potential victims from human trafficking.

The increasing use of technology by traffickers has prompted stakeholders to investigate how online platforms are employed to facilitate trafficking, leading to the identification of patterns, keywords, and indicators to help detect advertisements promoting sex trafficking victims (Hultgren et al., 2018; Dubrawski et al., 2015; Mayorga and Özaltın, 2023; Lugo-Graulich et al., 2024; Alvari et al., 2017; Giommoni and Ikwu, 2024). Furthermore, stakeholders are developing software solutions to combat human trafficking on online platforms (OSCE, 2020). Recent advancements like artificial intelligence (AI) have demonstrated significant potential for detecting human trafficking activity. Anti-trafficking software solutions have facilitated the analysis of numerous advertisements and have demonstrated significant efficiency as tools for identifying victims and supporting law enforcement and anti-trafficking organisations in their work (OSCE, 2020; Spotlight, n.d.; Marinus Analytics, 2022; Flores, 2022; Centric, 2022).

Most efforts to understand online human trafficking have focused on the exploitation stage after victims have already been recruited. Similarly, many technological solutions have focused on identifying trafficking activities during the exploitation stage, with most software tools focussing on detecting advertisements that promote victims of sex trafficking. There has been limited work on online recruitment (OSCE, 2020; Progetto Tenda et al., 2021; Thorn, 2018; Greiman and Bain, 2013; Ibanez and Suthers, 2014; Latonero, 2012; Zhou et al., 2024). The lack of attention to online recruitment can be attributed to various factors, such as data. Recruitment occurs predominantly on social media platforms, while online exploitation occurs primarily on public websites. Social media platforms face more restrictions, making data access more difficult than on public websites (Dorton, 2019; Latonero, 2012).

However, it is essential to find ways to gain a more comprehensive understanding of online recruitment, highlighting the methods that traffickers use to recruit victims and the circumstances that lead to their recruitment. The limited focus

on online recruitment has resulted in ineffective strategies to address this problem (Progetto Tenda et al., 2021; Zhou et al., 2024). Enhanced knowledge of these patterns and indicators will assist anti-trafficking stakeholders in developing more effective strategies to combat this crime. Similarly to technologies that detect sex trafficking advertisements, software tools can be developed to identify fraudulent job advertisements, enabling online platforms to remove them before potential victims are exposed.

Addressing recruitment can help prevent victims from experiencing traumatic events (Volodko et al., 2020). Neglecting this issue of online recruitment will only enable human traffickers to reach more vulnerable individuals on online platforms.

Although the development of software tools to address trafficking is important. It is equally critical to evaluate the risks of these software tools to ensure that their benefits are not outweighed by unintended negative consequences (Musto et al., 2020). It is crucial that these tools are developed with ethical considerations at the forefront to minimise harm, particularly to victims and marginalised communities.

## 1.1 The Research

This thesis aims to investigate human trafficking recruitment practices on digital platforms and to develop technological interventions to facilitate their detection while adhering to ethical considerations. It builds on previous research (Latonero and Movius, 2011; Mitchell and Boyd, 2014; Dubrawski et al., 2015; Mayorga and Özaltın, 2023; Alvari et al., 2017; Giommoni and Ikwu, 2024; Leary, 2014; Hultgren et al., 2016; Whitney, 2017; Hernandez-Alvarez, 2019; Simonson, 2021; Sarkar, 2015; Deeb-Swihart et al., 2019; Arenas et al., 2015) that focusses on understanding how victims are exploited online.

Furthermore, this research builds on existing work (Dorton, 2019; Latonero and Movius, 2011; Anthony, 2018; Dubrawski et al., 2015; Thorn, 2018; Latonero, 2012; Leary, 2014; Hultgren et al., 2016; Hernandez-Alvarez, 2019; Simonson, 2021; Sarkar, 2015; Arenas et al., 2015; Whitney et al., 2018; Tong et al., 2017; H. Wang et al., 2012; Silva et al., 2014; Alvari et al., 2016; Zhu et al., 2019; Wang et al.,

2020; Paula et al., 2023; Esfahani et al., 2020; J. Liu et al., 2023; Summers et al., 2023; Wiriyakun and Kurutach, 2021; Vajiac et al., 2022, 2023; Nair et al., 2022; Ardakani, 2020) that applies machine learning (ML) and natural language processing (NLP) to identify trafficking activity, extending the focus from exploitation to the recruitment stage. In addition, it introduces more recent technologies that have not yet been widely adopted in the field.

This research focusses on fraudulent job advertisements, as these advertisements are typically public data and represent one of the main recruitment methods. Analysing fake relationship recruitment is challenging because most interactions occur in chat rooms and there are limitations and privacy concerns around accessing such data.

This thesis also aims to incorporate measures to mitigate the potential harm from its research outputs. Given the lack of established measures in the anti-trafficking field (OSCE, 2020), this research aims to investigate the risks associated with anti-trafficking technologies and identify ways to mitigate them, contributing to the achievement of its research goals while also addressing calls for further research in this area (Latonero and Movius, 2011; Musto et al., 2020).

## 1.2 Research Question and Objectives

This thesis aims to address the following research questions and objectives.

### 1.2.1 Research Questions

- **RQ1:** What are the challenges and potential unintended consequences of developing and using anti-trafficking software, and how could these challenges and negative consequences be mitigated?
  - **RQ1.1** What are the design principles and guidelines that developers should consider when developing anti-trafficking software?
  - **RQ1.2** What guidelines should stakeholders consider when using anti-trafficking software?

  - **RQ1.3** How can the insights gained on the challenges and negative consequences of anti-trafficking software guide current research?

- **RQ2:** How do human traffickers recruit victims both offline and online, and what key factors and patterns contribute to this process?
- **RQ3:** What are the indicators of fraudulent job advertisements associated with human trafficking recruitment on online platforms, and how do anti-trafficking stakeholders effectively identify these indicators?
- **RQ4:** How can a set of AI models be developed to identify fraudulent job advertisements for online human trafficking recruitment while taking into account ethical considerations?

  **RQ4.1:** What procedures are necessary to collect and create a data set focused on fraudulent jobs associated with human trafficking?

### 1.2.2 Research Objectives

By addressing the research questions, the thesis aims to

- Identify the challenges and unintended consequences of developing and using anti-trafficking software and identify strategies to mitigate the challenges and negative unintended consequences observed.
- Gain insights into human trafficking recruitment in the real world and on online platforms.
- Identify indicators of fraudulent job advertisements for human trafficking on online platforms.
- Create a dataset dedicated to fraudulent job advertisements for human trafficking.
- Develop AI models capable of identifying fraudulent job advertisements related to human trafficking while taking into account the measures identified in the first objective.

- Evaluate the performance of the developed models using appropriate metrics and their effectiveness in helping various stakeholders identify this criminal activity.

## 1.3 Methodology

This thesis employs a mixed-method approach to address the research questions. Mixed Method Research (MMR) is a research process that integrates qualitative and quantitative techniques during data collection, analysis, and reporting (Shorten and Smith, 2017; Tashakkori and Creswell, 2007; Yu and Khazanchi, 2017). The researchers draw inferences from quantitative and qualitative data and benefit from different perspectives (Shorten and Smith, 2017; Creswell J. W. and Plano Clark, 2006). This approach is especially critical for research that cannot be effectively addressed using qualitative or quantitative methods alone.

This research incorporates the triangulation approach in MMR (Creswell J. W. and Plano Clark, 2006; O'Cathain et al., 2010). In the context of MMR, triangulation involves the use of multiple types of data and methods to gain a comprehensive understanding of research problems (Creswell J. W. and Plano Clark, 2006; O'Cathain et al., 2010). The data used are diverse, but complementary. The selection of MMR, specifically the triangulation approach, is well suited to the multidisciplinary nature of this thesis, as it seeks to investigate and detect the complex issue of online human trafficking recruitment while simultaneously aiming to mitigate potential negative effects. Moreover, the use of the MMR approach ensures the integration of expert feedback during the design of models and datasets, a recommendation made by several studies (Campana, 2022; Justice and Home Affairs Committee, 2022) when employing computational methods to combat human trafficking.

This study employs a multilevel research approach, which is one of several triangulation designs (Creswell J. W. and Plano Clark, 2006). This method involves collecting and analysing data at multiple levels to address a central research question (Creswell J. W. and Plano Clark, 2006). By integrating multiple research methods

into a single comprehensive study, this approach enables researchers to gain a complete understanding of the phenomenon under investigation (Creswell J. W. and Plano Clark, 2006). To effectively address the research questions, this thesis used a multilevel approach and is divided into four distinct phases, each with its data sources and methodologies. This approach was chosen because each research question required different types of data. The findings from each of the four phases informed one another, collectively contributing to the overall goal of the research.

Despite the advantages associated with the MMR approach, it also presents challenges, such as a steep technical learning curve (Shorten and Smith, 2017; Yu and Khazanchi, 2017; Creswell J. W. and Plano Clark, 2006). Researchers are expected to be proficient in both qualitative and quantitative methods, which can be challenging. Moreover, compared to the exclusive use of traditional qualitative or quantitative methods, many universities offer limited training in conducting mixed methods research (Yu and Khazanchi, 2017; Jick, 1979). To overcome these challenges, I devoted additional time to understanding all the methods used and proactively sought materials and online training resources related to MMR, both within and outside the university.

This thesis also aims to incorporate Responsible Research and Innovation (RRI) into the research process. RRI is an approach that involves examining the potential impacts of research and innovation while engaging with various stakeholders to ensure positive outcomes and minimise harm (European Commission, 2020; UK Research and Innovation, 2023; Owen et al., 2013; Jirotka et al., 2017). It helps identify potential risks to individuals and society that may arise from new technological developments and establishes measures to mitigate their negative effects.

This research develops a framework guided by the RRI methodology and integrates elements of this framework into various research activities to help minimise potential negative outcomes from its research outputs.

## 1.4 Contributions

This thesis offers valuable contributions to the field of computer science, while offering valuable insights and practical products to various stakeholders in the public and private sectors to address human trafficking recruitment. Examples of these stakeholders include law enforcement, anti-trafficking organisations, and policymakers.

The contributions include:

- Insights into the societal impact of anti-trafficking tools designed to identify victims and traffickers. These insights contribute to the creation of a framework that guides stakeholders in developing and deploying anti-trafficking software, helping minimise potential negative consequences. This work contributes to the body of research on the responsible and ethical use of technology in the fight against human trafficking.

- Provides a practical example of how to integrate RRI into studies that assess the potential implications of innovation.

- Improves understanding of how traffickers identify and recruit potential victims online and offline and presents design implications for technology that combats human trafficking recruitment. This information extends existing empirical insights in the field and helps with designing and creating more efficient computational and non-computational preventive measures.

- Highlights the indicators and patterns associated with fraudulent job advertisements linked to human trafficking on online platforms. It further examines the methodologies employed by anti-trafficking organisations to identify fraudulent job advertisements, offering insights that can inform software design. These findings assist stakeholders in identifying and mitigating risks associated with online recruitment while also providing critical considerations for platform design to prevent the inadvertent facilitation of trafficking activities. In

addition, they contribute to the development of technological solutions and datasets aimed at detecting fraudulent job listings.

- Presents a dataset annotated by experts, focussing on job offers associated with online human trafficking recruitment. This dataset provides the academic community with a valuable resource that can be used to conduct further research and develop technological solutions. In addition, this thesis also provides a larger machine-annotated dataset to aid in the creation of additional datasets to tackle this issue using AI advancements.

- Develops AI models that detect fraudulent job advertisements related to online human trafficking recruitment. These models will function as both performance and algorithmic benchmarks for detecting fraudulent job advertisements linked to human trafficking, providing baselines and a foundation for future research in the field of computer science. In addition, these models help evaluate the feasibility of using machine learning methods to identify fraudulent job postings online.

- Offers a valuable methodological contribution by incorporating design and development considerations aimed at mitigating negative effects during the creation of datasets and machine learning models for detecting fraudulent job postings. This approach provides a practical example of how the computer science field can integrate ethical considerations into developing anti-trafficking software.

  In addition, this thesis employs state-of-the-art artificial intelligence approaches that have not yet been applied in this field to develop models for online recruitment. It also explores the use of large language models to identify online recruitment, thereby expanding our understanding of the applicability and effectiveness of these models in combating human trafficking.

## 1.5 Terminology

The following terms are frequently referenced throughout the thesis:

- Technology-facilitated trafficking/ Online sex trafficking - This type of human trafficking involves the use of information and communication technologies (Latonero, 2012). In this thesis, "technology" or "Internet technologies" refer to software tools that enable users to exchange digital information.

- Anti-trafficking software (tools) - For this thesis, anti-trafficking tools are software tools that stakeholders use to assist them with combating human trafficking using various methods. This thesis will commonly refer to these tools as "tools" or "anti-trafficking tools" or "anti-trafficking software".

- Unintended consequences - This thesis adopts Franco-Santos and Otley's understanding of unintended consequences as "the outcome of a conscious action other than that foreseen and intended" (Franco-Santos and Otley, 2018).

- Stakeholders - These are people who use or develop anti-trafficking tools or are otherwise involved in combating human trafficking and supporting victims through various means.

- Online technologies/platforms - These terms primarily refer to Internet technologies, including social media platforms, websites, and communication applications.

- Job advertisements - These advertisements promote employment opportunities and provide details such as job description, title, qualifications, location, and application instructions. They are commonly disseminated through various channels such as newspapers and social media.

## 1.6 Publications

This thesis aims to produce three main publications, for which I will be the first author.

1. Moyo, T., Albayaydh, W., Webb, H., & Jirotka, M. (2025). A Responsible Research and Innovation Approach to Assessing the Implications of Anti-trafficking Technologies. Accepted to the Journal of Responsible Technology, ACM. This publication is based on Chapter 3.

2. Moyo, T., Gunes, O., & Jirotka, M. (2025). Investigating Human Trafficking Recruitment Online: A Study of Fraudulent Job Offers on Social Media Platforms. Accepted to CSCW 2025. This publication is based on Chapters 3 and 5.

3. Moyo, T., Gunes, O., & Jirotka, M. (2025). Benchmarking Traditional, Pre-trained Neural Language, and LLM-based Machine Learning Methods in Classifying Human Trafficking Job Advertisements. (In preparation). This publication will be based on Chapters 3, 4, 5, and 6.

## 1.7 Thesis Outline

Figure 1.1 provides a summary of the thesis. The structure of the thesis is as follows:

Chapter 2 provides background and context for the thesis. It explores the concept of human trafficking, the actors involved, and the stages of trafficking. This chapter also examines how technology has influenced and transformed the landscape of human trafficking. In addition, it reviews existing technological approaches used to identify and address human trafficking. Furthermore, the chapter highlights both technical and non-technical gaps in the existing literature, which inform the research questions.

Chapter 3 presents a qualitative study of the impact and challenges of developing anti-trafficking tools that identify victims or traffickers to combat sex trafficking. This chapter also provides recommendations on reducing the potential negative

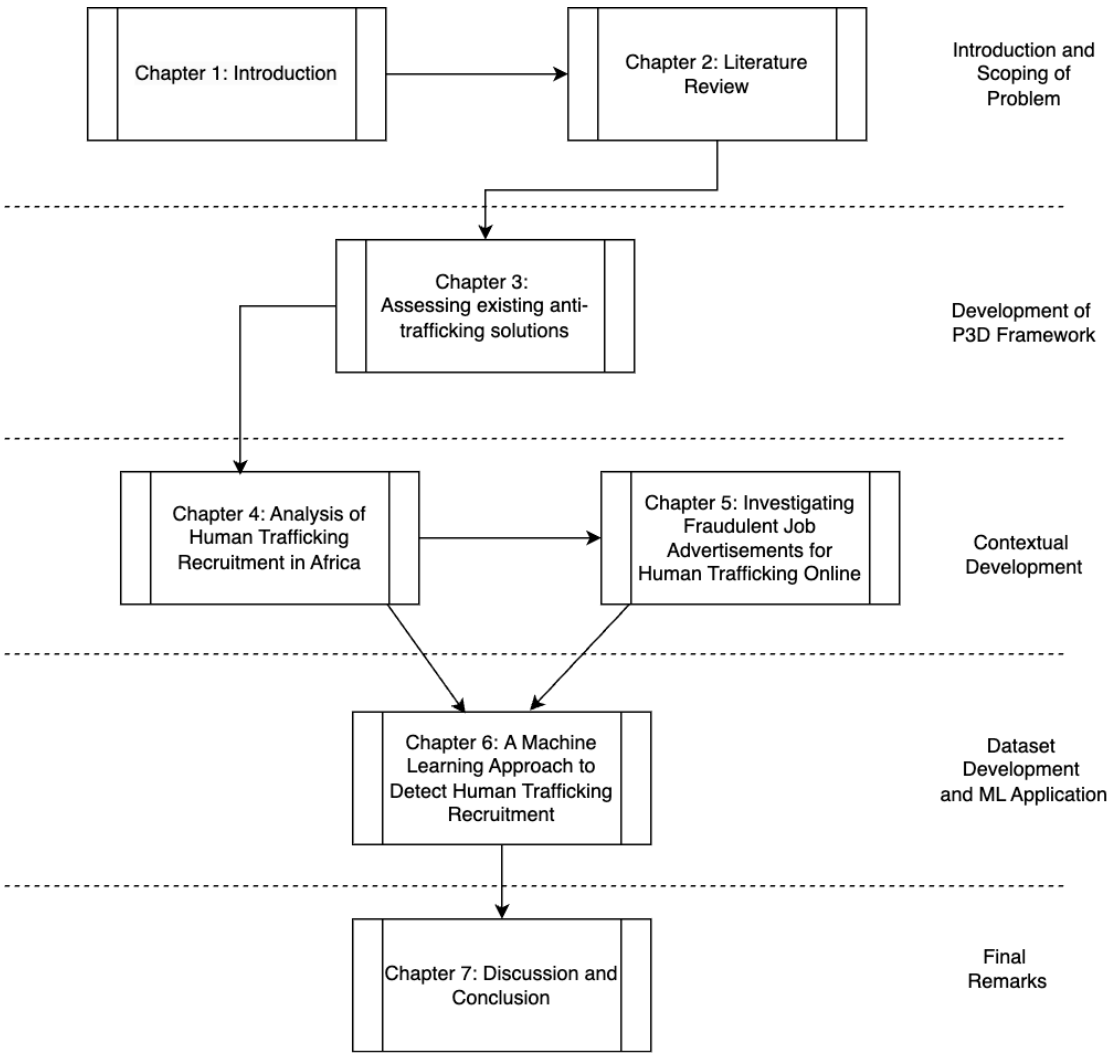


**Figure 1.1:** Summary of Thesis

consequences of these anti-trafficking tools. This study adopts a responsible research and innovation approach and demonstrates how the AREA framework was used to guide data collection and analysis.

Chapter 4 analyses two datasets related to the recruitment of victims of human trafficking across both online and offline platforms. It provides insights into recruitment methods, victims and traffickers' profiles, recruitment indicators, processes, and recurring trends, thereby offering a deeper understanding of human trafficking recruitment practices. This chapter also examines several hypotheses that explore fraudulent job offers in greater detail.

Chapter 5 presents an ethnographic study to acquire knowledge on how human

traffickers use social media to deceive people through fraudulent job advertisements. It also provides data on how anti-trafficking organisations identify and investigate fraudulent job opportunities on social media, as well as how they work to prevent potential victims from being recruited. Additionally, this chapter highlights the challenges faced by anti-trafficking organisations in identifying fraudulent job advertisements on social media.

Chapter 6 discusses the process of creating a novel public dataset based on job advertisements. It details the data collection and labelling process. In addition, this chapter compares the performance of traditional machine learning models, pre-trained neural language models (PLMs), and large language models (LLMs). The most effective approach among the models is used to create a machine-labelled dataset.

Finally, Chapter 7 discusses the findings and contributions of this thesis. Furthermore, this chapter sheds light on the challenges faced in developing anti-trafficking tools in this field and provides recommendations to various stakeholders.

# 2

# Background and Literature Review

## Contents


**2.1 Introduction . . . . . 16**
**2.2 Overview of Human Trafficking . . . . . 17**
2.2.1 Types of Human Trafficking . . . . . 17
2.2.2 Human Trafficking Actors . . . . . 19
2.2.3 Stages of Human Trafficking . . . . . 22
**2.3 Overview of Technology-Facilitated Trafficking . . . . . 24**
2.3.1 Online Recruitment . . . . . 26
2.3.2 Online Exploitation . . . . . 31
**2.4 Anti-Trafficking Tools . . . . . 34**
**2.5 The Impact of Anti-Trafficking Tools . . . . . 37**
**2.6 Related Work on Technological Solutions to Detect Online Trafficking . . . . . 39**
2.6.1 Data Sources and Acquisition . . . . . 40
2.6.2 Detecting Interned-based Trafficking using Machine Learning and Natural Language Processing . . . . . 42
**2.7 Anti-trafficking Policies and Online Legislation . . . . 49**
**2.8 Summary and Current Limitations . . . . . 51**
2.8.1 Limited Studies that Assess the Impact of Anti-Trafficking Tools . . . . . 51
2.8.2 Limited Work on Online Recruitment . . . . . 52
2.8.3 Lack of Anti-trafficking Technologies that Identify Online Recruitment . . . . . 53
2.8.4 Limited Use of State-of-the-art Methods in Natural Language Processing . . . . . 54
**2.9 Direction of Thesis . . . . . 55**

## 2.1 Introduction

This chapter provides a context for this thesis, focussing on technology-facilitated trafficking, and highlights the technical and non-technical gaps in this domain. To identify relevant literature, a variety of searches were performed across various platforms, primarily Google Scholar and SOLO, the university library catalogue. This catalogue includes non-print legal deposit books, scores, journal articles, and issues, as well as the UK Web Archive. The searches were not restricted to any particular repository, allowing access to papers from any source containing relevant information on anti-trafficking tools.

Several keywords were employed during the search, including "sex trafficking", "human trafficking", "technology-facilitated trafficking", "online trafficking", "human trafficking Internet", "technology trafficking", "escort ads online human trafficking", "detecting human trafficking", "detecting sex trafficking", "online sex trafficking", "anti-trafficking", and "child sex trafficking". The combination of keywords provided coverage of the literature on human trafficking, technology-facilitated trafficking, and anti-trafficking efforts. These keywords help identify both general and specific areas within the research scope.

In addition to database searches, snowball sampling was used to identify other relevant papers. This method involved examining the reference sections of the reviewed articles to discover additional studies.

This chapter is organised as follows. Firstly, Section 2.2 examines human trafficking and discusses its stages, actors, and types. Section 2.3 addresses technology-facilitated trafficking and analyses the use of technology in victim recruitment and exploitation. Sections 2.4 and 2.5 discuss anti-trafficking tools and their effectiveness. Section 2.6 examines the computational approaches used in human trafficking detection with a focus on academic research. Section 2.7 discusses existing legislative frameworks that address online human trafficking. Section 2.8 summarises the research gaps identified in the review of the literature.

Finally, Section 2.9 discusses how the literature review shaped the thesis direction in terms of the type of human trafficking, online recruitment methods, and the location it focuses on.

## 2.2 Overview of Human Trafficking

The United Nations Protocol on Trafficking in Persons, signed by member states to address modern-day slavery, defines human trafficking as:

> "The recruitment, transportation, transfer, harbouring, or receipt of persons, by means of the threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of a person having control over another person, for the purpose of exploitation. Exploitation shall include, at a minimum, the exploitation of the prostitution of others or other forms of sexual exploitation, forced labour or services, slavery or practices similar to slavery, servitude, or the removal of organs." (United Nations, 2000)

The protocol further stipulates that any recruitment, transportation, harbouring, or exploitation involving minors will automatically be classified as human trafficking, regardless of the conditions mentioned above (United Nations, 2000). Criminal networks involved in human trafficking operations have transformed this illicit activity into profitable enterprises, exploiting victims across international borders with considerable ease (OSCE, 2020; Allais, 2006). Human trafficking is highly lucrative because of its low prosecution rates, insufficient legislation, and widespread corruption (OSCE, 2020; Allais, 2006). These factors contribute to the continued prevalence of human trafficking around the world.

### 2.2.1 Types of Human Trafficking

There are various types of human trafficking, and the following section discusses these types in more detail.

### Sex Trafficking

The U.S. Victims of Trafficking and Violence Protection Act (TVPA) of 2000 defines sex trafficking as recruiting, harbouring, transporting, providing, or obtaining a person with the intent of engaging them in commercial sexual activities (U.S. Congress, 2000). Sex trafficking and sexual exploitation will be used interchangeably in the thesis.

The majority of victims detected worldwide are for sex trafficking (United Nations Office on Drugs and Crime, 2023, 2020, 2018); however, this number has decreased during the COVID-19 pandemic (United Nations Office on Drugs and Crime, 2023). Traffickers force victims to work as prostitutes or perform other sexual activities (Crawford, 2017).

Sex trafficking is common in regions such as America, Europe, East Asia, and the Pacific (United Nations Office on Drugs and Crime, 2018). Europe, America, and parts of Asia report the highest number of detected victims, which explains why the majority of identified victims are women recruited for sexual exploitation (United Nations Office on Drugs and Crime, 2018). Some countries, particularly developing nations, lack adequate measures to detect victims, leading to fewer victims being identified (United Nations Office on Drugs and Crime, 2018).

### Forced Labour

The International Labour Organisation defines forced labour as "All work or service which is demanded from any person under the menace of any penalty and for which the said person has not offered himself voluntarily." (International Labour Office. Special Action Programme to Combat Forced Labour., 2008). Forced labour applies to adults and children and accounts for all forms of employment. Voluntary work allows individuals to enter and leave work at their own risk (International Labour Office. Special Action Programme to Combat Forced Labour., 2008).

For many years, sexual exploitation has been detected in significantly higher numbers than forced labour; however, in 2020, the numbers of forced labour and sexual exploitation were equal (United Nations: Office on Drugs and Crime, 2023).

This increase was gradual, with a slight decrease between 2019 and 2020. The men constituted 68%, while the women constituted 32% of the detected victims, and the majority of the victims were recruited for the food and food supply chain (United Nations: Office on Drugs and Crime, 2023).

#### Other Forms of Exploitation

Sex trafficking and forced labour are the primary forms of exploitation, and the rest of exploitations are grouped as "other forms of exploitation" (United Nations: Office on Drugs and Crime, 2023). Examples of other forms of exploitation include organ harvesting, forced marriage, forced adoption, and forced criminality (pickpocketing, begging, drug trafficking, and fraud). In 2020, 2% of the detected victims were detected for exploitative begging and forced marriage (1% each). Other forms of human trafficking are more difficult to detect because of significant variations in legal status between different countries (United Nations Office on Drugs and Crime, 2018).

United Nations Office on Drugs and Organised Crime (UNDOC) report states that victims involved in other types of exploitation are commonly detected in circumstances involving sex trafficking and forced labour (United Nations: Office on Drugs and Crime, 2023).

### 2.2.2 Human Trafficking Actors

A 2020 report by the Organisation for Security and Co-operation in Europe (OSCE), an organisation that enhances security and maintains peace and stability in its member states that span Europe, Central Asia and North America, states that human trafficking exhibits characteristics similar to a marketplace in which individuals engage in the trade and commodification of victims (OSCE, 2020). There is demand, supply, competition, and pricing (Whitney, 2017). Victims are considered supply items for economic gains, while buyers represent demand (Mitchell and Boyd, 2014). There is competition between traffickers in terms of prices. The primary goal of human trafficking is to maximise profits by exploiting more victims (OSCE, 2020; Arronte, 2018). This commodification of victims as part of an economic transaction

reduces them to objects of profit rather than recognising them as individuals with rights. This treatment further perpetuates their mistreatment. The following section provides further details on victims and perpetrators of human trafficking.

### Trafficking Victims

Human traffickers target people of both genders; however, women constitute the majority of victims. In 2020, women comprised 60% of the identified victims, while men represented 40% among the identified victims (United Nations: Office on Drugs and Crime, 2023). Over time, the proportion of female victims detected has declined, whereas the number of male victims detected has increased. This shift may be attributed to the reduction in sex trafficking and the concurrent rise in forced labour, which predominantly affects male victims.

Several factors increase the vulnerability of individuals to human trafficking, such as unstable domestic environments, poverty, conflicts, wars, illegal immigration status, and substance abuse (Thorn, 2018; United Nations: Office on Drugs and Crime, 2023; United Nations Office on Drugs and Crime, 2012). Human traffickers exploit these vulnerabilities to trap victims. The following section examines different types of human traffickers and their strategies.

### Human Traffickers

Traffickers can be family, friends, employers, or strangers (United Nations Office on Drugs and Crime, 2012; US Department of State, 2019; UNICEF, 2003). This paper will categorise traffickers into first-party traffickers (buyers) and third-party traffickers (suppliers).

**Third-Party Traffickers-** Third-party traffickers are individuals who engage in the recruitment, coercion, transportation, control, and exploitation of people (Leary, 2014; Arronte, 2018). These types of traffickers are also known as pimps, exploiters, or controllers, as they typically employ physical force or intimidation to compel victims into exploitation (Leary, 2014; Davis, 2013; Mitchell et al.,

2011). Third-party traffickers can assume the role of one or more of the following categories to gain victims' trust (Riley, 2016):

- Pretender – These individuals present a false front to deceive victims (Anthony, 2018). For example, family members or friends may exploit victims under the pretense of care and affection.
- Provider – These are individuals who provide victims with their needs, both necessities and luxury items. Examples include buying them clothes and phones or paying their rent.
- Promiser - These individuals lure victims with false promises of a better life by promising them things such as a good job, fame, or access to a glamorous lifestyle (Anthony, 2018; US Department of State, 2019).
- Protector – These individuals use physical power and intimidation to control victims. The victim views the trafficker as their protector when they form a trauma bond with them. A trauma bond is only created when victims cannot escape and have a death threat, feel isolated, or perceive a form of kindness from the trafficker (Fonrouge, 2018).
- Punisher – Traffickers typically use this role as a last resort to control victims after exhausting the other roles. They use physical violence, threats, and blackmail to control victims (Anthony, 2018; Sarkar, 2015; US Department of State, 2019; Konrad et al., 2016).

The roles are not mutually independent; traffickers can assume multiple roles or switch between them depending on the circumstances.

The primary objective of third-party traffickers is to exploit victims for financial gain (OSCE, 2020; Arronte, 2018). Traffickers may operate independently or as part of a larger criminal organisation (Mitchell and Boyd, 2014). Organised criminal groups in human trafficking consist of at least three individuals who collaborate to exploit people for financial or other gains (United Nations: Office on Drugs and Crime, 2023). Sarkar states that third-party traffickers have connections with

influential figures, including politicians, state officials, police officers, and border security personnel (Sarkar, 2015). These connections can make disrupting traffickers particularly challenging. Preventive interventions against victims may be more effective in dismantling their operations. This thesis primarily focusses on this category of traffickers and henceforth refers to them simply as "traffickers" or "human traffickers".

**First-Party Traffickers-** First-party traffickers are individuals who typically exploit individuals for their own gratification and not for financial gain or profit (Sykiotou, 2017). These individuals can be clients in the supply chain (Sykiotou, 2007). In sex trafficking, any individual who engages in or attempts sexual activities with a minor, whether coerced or not, is classified as a trafficker (Latonero and Movius, 2011; US Department of State, 2020). A child sex trafficker can be a supplier of children or a consumer of commercial sexual services. The majority of first-party traffickers in sex trafficking are travelling business professionals seeking local prostitution services (Sykiotou, 2007; Sarkar, 2015; Sykiotou, 2017). The purchase of services from a trafficked individual also constitutes a criminal offence in various regions of the world, such as Europe (Sykiotou, 2007). This thesis refers to these types of traffickers as "first-party traffickers" or "buyers". The next section examines some of the stages involved in human trafficking.

### 2.2.3 Stages of Human Trafficking

The standard stages of human trafficking include recruitment, transportation, and exploitation.

During the recruitment stage, the traffickers use various methods to target vulnerable individuals. Traffickers use coercion, force, or fraud to recruit victims (Newsome Melton, 2020). Coercion refers to the use of threats, whereas fraud involves the dissemination of incorrect information to victims with the intent of deceiving them (Anthony, 2018; Olivette, 2021). Force encompasses physical

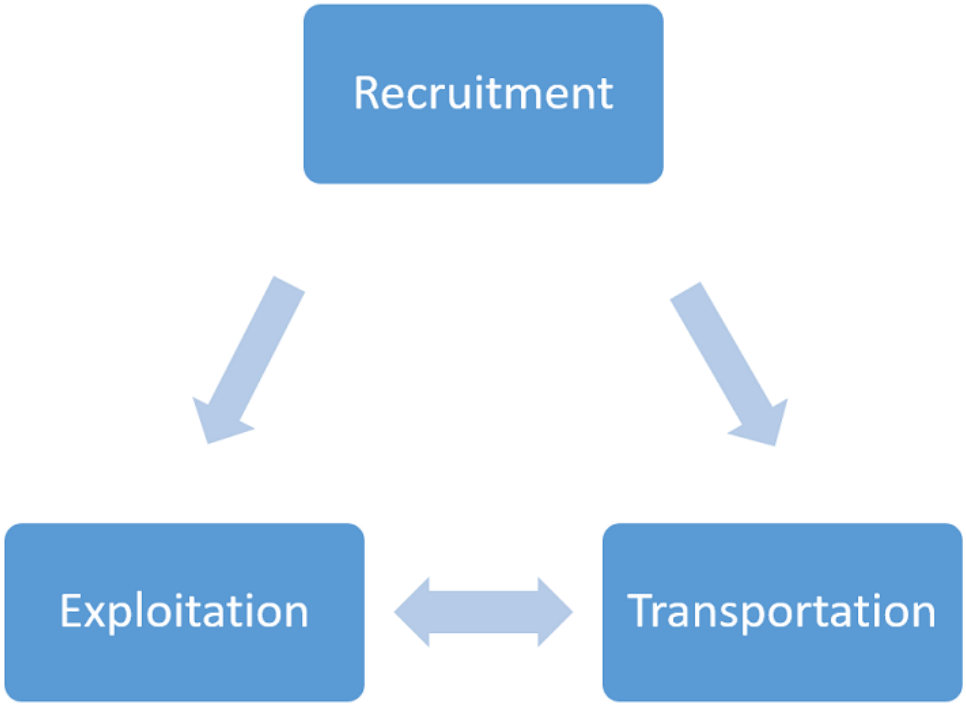


**Figure 2.1:** The Stages of Human Trafficking

violence or forms of psychological manipulation such as compulsion, constraint, and restraint (US Department of State, 2019; Olivette, 2021).

The transportation stage encompasses the relocation of victims from one geographical location to another using various modes of transport, such as roads, seas, and air (Olivette, 2021). Victims may be trafficked individually or as part of a larger group. Traffickers tend to transport victims within their country of origin or across international borders against their will (United Nations: Office on Drugs and Crime, 2023). Human trafficking can occur in the absence of the transportation stage (US Department of State, 2019). The act of transporting victims is not a prerequisite for a situation to be classified as human trafficking.

During the exploitation phase, traffickers force victims to engage in activities against their will, including various forms of abuse, such as forced labour and prostitution (Olivette, 2021). Traffickers use a variety of coercive strategies to ensure compliance, including physical restraint, psychological manipulation, and isolation from family and social support networks (Newsome Melton, 2020).

Regardless of the form of exploitation, the consequences for victims are severe

and affect their physical, mental, and emotional well-being. Even in survival cases, victims' reintegration into society and resumption of normal life can present significant challenges. Therefore, substantial efforts must be made to prevent exploitation from happening in the first place.

Technology has significantly transformed the way human traffickers recruit and exploit individuals. The following sections provide further details on how human traffickers use technology in their operations.

## 2.3 Overview of Technology-Facilitated Trafficking

Human trafficking occurs predominantly in the physical environment. However, in recent years, as previously mentioned, traffickers have begun to use various technological platforms, particularly the Internet, to facilitate trafficking activities (Latonero and Movius, 2011; Sykiotou, 2007; Arronte, 2018; Sykiotou, 2017). Online platforms prove effective due to the traffickers' ability to access individuals from diverse geographical areas, thereby recruiting and exploiting a larger number of individuals than would typically be possible. Several terms have been used to describe this phenomenon, including internet-based trafficking, online trafficking, cyber-trafficking, technology-facilitated trafficking, and e-trafficking. These designations primarily refer to the use of Internet technologies such as social media platforms, websites, and communication applications to facilitate trafficking operations. This thesis will focus on these forms of trafficking and refer to them as "technology-facilitated trafficking" or "online trafficking".

Human traffickers use various existing Internet technologies (Dorton, 2019; OSCE, 2020; Latonero and Movius, 2011; United Nations Office on Drugs and Crime, 2018; US Department of State, 2019), and in some cases, they establish their own sophisticated websites or hire individuals to develop these platforms. Custom platforms are often tailored to target specific demographics for recruitment and to cater to their clientele (United Nations Office on Drugs and Crime, 2020; Sarkar, 2015; Sykiotou, 2017). The developers of these websites aim to ensure

that the traffickers conceal electronic traces from law enforcement and operate anonymously (Sykiotou, 2007; Mitchell and Boyd, 2014; Sarkar, 2015; Gerry QC et al., 2016). The acquisition of data for investigations is challenging, particularly when traffickers use protective technologies to conceal their activities. The act of creating their websites or hiring individuals to do so highlights the growing reliance of traffickers on technology.

The thesis mainly discusses four types of online platforms:

1. Online classified websites – These online platforms facilitate rapid online transactions for buying and selling goods and sometimes incorporate sections for anonymous postings of services and personal advertisements (OSCE, 2020; Latonero, 2012). Some of these classified online platforms feature adult sections that advertise commercial sexual services. Although these websites may appear harmless, they have become focal points for sex trafficking activities (C. Rodriguez, 2018). Examples of such websites include Craigslist and Backpage.

2. Adult/Escort websites – These websites facilitate connections between clients and individuals engaged in sex-related work. Although the primary purpose of escort websites is to connect consenting adults, these platforms also function as online communities where sex traffickers exploit victims and facilitate their interactions with clients (Mitchell and Boyd, 2014; Leary, 2014). Examples of such websites include Cosmos Escorts International and International Escort Agency (Sykiotou, 2007).

3. Social media platforms. - These platforms have gained significant popularity, with millions of users using them to establish connections, exchange messages, post content including images and videos, and expand their networks (Latonero and Movius, 2011; Sarkar, 2015; Siddiqui, 2016; Arli, 2017). Personal data and other online information shared on social media contribute to an individual's digital footprint, which cybercriminals exploit to target potential victims (Arronte, 2018). Examples include Twitter, Facebook, and Instagram (Dorton,

2019; Anthony, 2018). This thesis employs the terms "social networking sites" and "social media" interchangeably.

4. Job boards - These online platforms function as search engines for employment opportunities, facilitating the advertisement and exploration of job vacancies (Petrosyan, 2022). These websites have gained significant popularity, serving as intermediaries between employers and prospective employees. Human traffickers exploit these platforms to recruit people through fraudulent job advertisements. Prominent examples of job boards include Monster and Indeed.

The following sections will discuss how these technologies are used for recruitment and exploitation.

### 2.3.1 Online Recruitment

Online recruitment refers to the process by which human traffickers use online platforms to identify and recruit victims. In recent years, traffickers have demonstrated an increased tendency to recruit victims online rather than through traditional offline methods (OSCE, 2020; Department of State, 2023; Feehs and Wheeler, 2020; Europol, 2014; Ibanez and Suthers, 2014; Walby et al., 2016). Traffickers prefer online platforms for recruitment due to their ability to evade the risk of detection by law enforcement, attributed to the numerous features inherent in these platforms and the easy access to information that can exploited for recruitment (OSCE, 2020; Progetto Tenda et al., 2021; Sykiotou, 2007). Although online recruitment is frequently perceived as more prevalent in developed countries due to the widespread adoption of the Internet, research has shown that traffickers also use online platforms to recruit victims in developing countries and rural areas (Anthony, 2018). However, most of the literature has focused on online recruitment in western countries.

The issue of online recruitment predates the widespread adoption of the Internet, with research showing that victims were recruited through MySpace in the early 2000s (Anthony, 2018). In recent years, social media platforms and various types of

websites, such as job boards and marriage websites, are among the recent online platforms used by traffickers to conduct recruitment (Dorton, 2019; Latonero and Movius, 2011; Feehs and Wheeler, 2020; Sarkar, 2015; Arronte, 2018; Walby et al., 2016; Allen, 2019). Facebook has emerged as one of the primary online platforms for traffickers, as the majority of cases involving online recruitment occur on this platform (Dorton, 2019; Progetto Tenda et al., 2021; Feehs and Wheeler, 2020). Facebook enables users to advertise products and services, post videos and pictures, and connect through messaging. In recent years, it has consistently maintained the highest number of social media users (Dorton, 2019). In the last quarter of 2023, Facebook had an estimated 3.065 billion active users on its platform (Shepherd, 2024). This prevalence may explain why Facebook is a preferred platform for traffickers. Other popular platforms such as WeChat and Instagram follow closely behind Facebook (Progetto Tenda et al., 2021; Feehs and Wheeler, 2020). A survivor of sex trafficking states that:

> "[As a survivor advocate, I've seen] a huge influx [of social media recruitment] now... I was recruited through a dating website. I don't think they had direct messaging on social media platforms back [when I was recruited]. But a lot of my clients are recruited off Facebook or Instagram. [Recruiters] send them direct messages... I have girls who are flown from all over the country... and they think they are coming to see a modelling agent" (Anthony, 2018).

The COVID-19 pandemic exacerbated the issue of online recruitment, as traffickers increasingly used online platforms to recruit victims (Progetto Tenda et al., 2021). Children exhibited heightened vulnerability to online grooming for sex trafficking because of the substantial amount of time they spent online during the pandemic. A 2020 federal report by the Human Trafficking Institute indicated that online recruitment was the predominant recruitment tactic used by traffickers, according to active cases under prosecution (Feehs and Wheeler, 2020).

Traffickers implement various strategies to recruit victims online. The primary recruitment strategies can be classified as "hunting" and "fishing" (United Nations Office on Drugs and Crime, 2020). The hunting strategy involves traffickers targeting

specific individuals based on their vulnerabilities, such as economic or emotional difficulties, and those who are likely to be susceptible to exploitation (United Nations Office on Drugs and Crime, 2020). These perpetrators identify potential victims by analysing personal information shared on social media platforms (Department of State, 2023; Progetto Tenda et al., 2021; United Nations Office on Drugs and Crime, 2020; Arronte, 2018). However, a clear profile of the people targeted online remains unclear, preventing the development of effective interventions (OSCE, 2020). More research is needed to understand the characteristics of victims of technology-facilitated trafficking. Victim profiles may vary by country or region, making it essential for each region to understand these profiles in order to develop targeted measures. For example, in EU member states, the majority of victims targeted by online recruitment were adult women (Europol, 2020); however, this may not be generalisable to other regions.

The fishing strategy involves victims initiating contact due to fraudulent information disseminated online by traffickers (Progetto Tenda et al., 2021; Ramchandani et al., 2021; United Nations Office on Drugs and Crime, 2020). Victims identified through this strategy are random, as they are contingent upon who responds to the information posted online. Hunting strategies target victims primarily on social media platforms (Progetto Tenda et al., 2021; United Nations Office on Drugs and Crime, 2020; Sarkar, 2015), while fishing strategies are generally carried out on classified websites, job boards, social media, and marriage websites (Anthony, 2018; Sykiotou, 2007).

As previously mentioned, there are two main methods that traffickers use to recruit their victims online: advertising fraudulent jobs and establishing fake online relationships in different countries, such as the US and Italy (Progetto Tenda et al., 2021; Anthony, 2018; Feehs and Wheeler, 2020). Fake online relations involve human traffickers establishing a fraudulent relationship with potential victims, exploiting their vulnerabilities based on information obtained from their profiles (Anthony, 2018). The process typically begins with the collection and analysis of a victim's personal information to cultivate a relationship through social media

platforms (Progetto Tenda et al., 2021; Anthony, 2018; United Nations Office on Drugs and Crime, 2020). Subsequently, traffickers develop rapport and intimacy with victims using strategies such as gift giving and romantic correspondence, which ultimately lead to transportation (Anthony, 2018). The following section provides a more detailed analysis of the primary method addressed in this thesis: fraudulent job advertisements.

### Fraudulent Job Advertisements

Job advertisements have become a prevalent method of recruiting people. These advertisements frequently employ persuasive language, promising potential victims, an improved lifestyle, and offer substantial incentives (Anthony, 2018; United Nations Office on Drugs and Crime, 2020). Upon receiving responses, traffickers persuade victims to enter into contractual agreements and in cases involving international travel, work documents are processed, which can result in debt accumulation (Department of State, 2023; Sarkar, 2015). This debt is used as a control mechanism, leading to exploitation, as the majority of the victims are unable to repay it. In certain instances, traffickers resort to coercion or physical violence after establishing initial contact with victims (Progetto Tenda et al., 2021; Anthony, 2018).

For example, a group of traffickers established fictitious profiles on social media platforms and advertised fraudulent employment opportunities on the pages they created (United Nations Office on Drugs and Crime, 2018). They successfully recruited 100 girls, coercing them to send compromising photographs as a means of control. In 2018, a human trafficking survivor initiated legal proceedings against Facebook, alleging that the platform facilitated traffickers to recruit, and recruit children into the sex trade (Freedom United, 2018). This allegation claimed that Facebook offered an accessible platform for sex traffickers to contact minors. This victim was lured into sex trafficking at age 15 through a job offer, and endured various types of physical abuse. These cases highlight the critical need to identify and remove fraudulent job advertisements before traffickers establish contact with potential victims and coerce them to be exploited. A 2023 US Department of

State report states that the prevalence of fraudulent job advertisements online presents a significant challenge for victims in distinguishing between legitimate and illegitimate job opportunities (Department of State, 2023).

Victims are frequently recruited by unidentified individuals who conceal their true identities (Anthony, 2018; Arronte, 2018). Research indicates (Xiao et al., 2015) that numerous fraudulent profiles on social media platforms enable traffickers to operate anonymously (Arronte, 2018). However, despite this anonymity, traffickers leave behind evidence that can be used for detection, and measures can be implemented to prevent them from disseminating fraudulent job offers. Simonson proposed the development of a software tool to identify fraudulent job advertisements and remove them before they reach potential victims (Simonson, 2021).

To effectively detect and remove fraudulent job offers before their dissemination to potential victims, a more comprehensive understanding of the methods used to disseminate these job advertisements and their content is needed. A previous study identified indicators of human trafficking in job advertisements on a Lithuanian website, focusing on labour trafficking (Volodko et al., 2020). This study used offline indicators based on relevant legislation to assess their effectiveness in detecting trafficking. It successfully identified offline indicators of human trafficking in these job advertisements (Volodko et al., 2020). However, it is necessary to develop indicators specifically tailored for technology-facilitated trafficking rather than relying solely on offline indicators (Campana, 2022). When these indicators are identified, similar research can be conducted to explore the prevalence of online indicators of fraudulent job posts across various online platforms.

One of the challenges in addressing this issue is the diversity of online platforms used by traffickers to disseminate fraudulent job advertisements. These platforms are selected based on their geographical location, which impedes stakeholders' comprehension of their practices because of the extensive range of platforms involved (Department of State, 2023). It is necessary for stakeholders to focus on identifying the prevalent platforms in various regions and improve their understanding of the specific platforms to develop targeted strategies to remove fraudulent job

offers. More individual analyses of different platforms will lead to an overall improved understanding of the issue.

For more than a decade, stakeholders have advocated greater efforts to understand online recruitment methods (Department of State, 2023; Thorn, 2018; Greiman and Bain, 2013). However, comprehensive information on the intricacies of online human trafficking recruitment remains insufficient, including fraudulent job advertisements (OSCE, 2020; Progetto Tenda et al., 2021; Thorn, 2018; Greiman and Bain, 2013; Ibanez and Suthers, 2014; Latonero, 2012; Zhou et al., 2024). Current research has not thoroughly examined the indicators and characteristics of these recruitment methods. This information would facilitate the development of a list of indicators for online recruitment and contribute to the creation of solutions to detect recruitment activities. The need for more comprehensive information and understanding of online recruitment has prevented efficient means of identifying recruitment patterns on the Internet (Progetto Tenda et al., 2021; Zhou et al., 2024), further highlighting more work in this area.

Following the recruitment of victims, the traffickers engage in exploitation through various online platforms. The next section analyses how victims are exploited online.

### 2.3.2 Online Exploitation

Human traffickers use online platforms to advertise and sell victims of various forms of exploitation, particularly sex trafficking (Europol, 2014; Mitchell and Boyd, 2014; Thorn, 2018; Sarkar, 2015). Online catalogues containing photographs of women and girls, along with their contact details, are used to promote and facilitate prostitution for commercial sex and other sexual activities (United Nations Office on Drugs and Crime, 2020; Europol, 2014; Mitchell and Boyd, 2014; Sarkar, 2015). Potential buyers search through these online catalogues to seek the services of victims (Sykiotou, 2007, 2017).

A 2018 report by Thorn, a US-based organisation that develops software tools to combat sex trafficking, indicates that recent victims of sex trafficking are more

likely to be involved in online prostitution than street prostitution (Thorn, 2018). Online prostitution is linked to a higher number of buyers, as it enables connections between victims and buyers across multiple geographical locations, rather than being confined to the victims' immediate vicinity. Furthermore, a 2020 study by the Human Trafficking Institute, an organisation that combats trafficking through strategic legal frameworks and systematic efforts in the US, found that 83% of active sex trafficking cases in the past 20 years involved the Internet as a solicitation tool (Feehs and Wheeler, 2020). Similar studies in Europe, such as in Italy, have also found that most of the current cases involve online trafficking (Feehs and Wheeler, 2020).

Online technologies have significantly facilitated the production and distribution of CSAM (Child Sexual Abuse Material), leading to the exploitation of millions of children online (Lockett, 2018; Mitchell and Boyd, 2014). Evidence suggests that traffickers coerce children as young as five years to commit sexual acts using webcams to reach a global audience (Lockett, 2018). In the Philippines, the number of reported cases of cybersex involving children ranges from 60,000 (Blomberg, 2019) to one million (Lockett, 2018). Furthermore, law enforcement agencies arrested traffickers and buyers for the online sale of 385 infants as part of private adoption (Sarkar, 2015).

Victims are advertised on online platforms such as classified sites, adult websites, and social media platforms (Latonero and Movius, 2011; Latonero, 2012; Arronte, 2018). Adult websites such as Adult Search and My Redbook have more than 100,000 unique users per month (Latonero, 2012). Classified websites, such as Craigslist and Backpage, were standard websites used by traffickers to market women and girls before law enforcement shut down these websites (Dorton, 2019). Backpage is a website that advertises women and girls for commercial sex. They requested that traffickers edit advertisements to make them less noticeable to law enforcement (Anthony, 2018). In 2013, Backpage was responsible for 80% of the online sex market, from advertising to the sale of commercial sex. Eventually, law enforcement shut down Backpage, along with an adult section on Craigslist that facilitated sex trafficking, but this only resulted in traffickers using other social media technologies to advertise and sell women and girls (Dorton, 2019).

A prevalent strategy used by traffickers involves using victims' personal social media accounts to attract potential clients (Dorton, 2019; Mitchell and Boyd, 2014; Sarkar, 2015). Previous research (Thorn, 2018) revealed that the majority of interviewed victims of child sex trafficking reported creating their own advertisements using specific terminology. However, despite the significant presence of human trafficking on social media, most stakeholders focus on monitoring adults and classified websites because of the challenges associated with social media. For example, law enforcement agencies have indicated that they typically monitor adult websites to identify potential trafficking cases, but do not generally survey social media platforms because of the increased difficulty in detecting trafficking activities on those platforms compared to adult websites (Latonero, 2012). Furthermore, much of the research on online exploitation has focused on classified or escort websites rather than social media platforms (Dorton, 2019). Data from classified and adult websites are generally more accessible, whereas data from social media present greater challenges for investigation. Most social media platforms have strict rules regarding data collection, which can have legal consequences. More efforts should be put into establishing partnerships with these companies to develop solutions to combat human trafficking.

Most studies on technology-facilitated trafficking have focused on understanding the exploitation stage (Latonero and Movius, 2011; Mitchell and Boyd, 2014; Dubrawski et al., 2015; Mayorga and Özaltın, 2023; Alvari et al., 2017; Giommoni and Ikwu, 2024; Leary, 2014; Hultgren et al., 2016; Whitney, 2017; Hernandez-Alvarez, 2019; Simonson, 2021; Sarkar, 2015; Deeb-Swihart et al., 2019; Arenas et al., 2015; Nair et al., 2022). These studies have investigated how traffickers advertise and sell victims online and how law enforcement investigates such cases. Several indicators used by different stakeholders to detect online exploitation have been identified. Examples include keywords such as "new in town", "turned 18", "daddy's little girl", and "youthful" (Hultgren et al., 2018; Dubrawski et al., 2015; Mayorga and Özaltın, 2023; Lugo-Graulich et al., 2024; Alvari et al., 2017; Giommoni and Ikwu, 2024), as well as a list of emojis, such as roses, cherries, cherry blossoms,

and growing heart have been identified as being used to communicate information about human trafficking victims between first- and third-party traffickers in online environments (Whitney, 2017; Whitney et al., 2018). Examples of terms frequently associated with illicit activities with massage parlours include "happy ending", "prostitution", and "tableshower". Illicit massage parlors are businesses that exploit people for forced labour and sex trafficking (Mayorga and Özaltın, 2023; Li et al., 2023). The significant body of research on online exploitation further highlights the disparity between the attention given to online recruitment and exploitation.

Various automated solutions have been developed to detect these trafficking activities during the recruitment and exploitation stage. The following section discusses these technologies in more detail.

## 2.4 Anti-Trafficking Tools

In response to the increasing use of technology, several software solutions have been developed to combat trafficking. Traffickers and victims often leave critical information in the online content they generate that can potentially facilitate their identification. An OSCE report indicates that anti-trafficking tools fall into various categories, including Victim/Trafficker Identification, Awareness Raising, Education and Collaboration, Supply Chain Management, Data Trends and Mapping, Corporate Risk Identification, Worker Engagement and Empowerment, Victim Case Management and Support, Ethical shopping, Privacy/Personal data, Payment security, and others (OSCE, 2020). The most prevalent tools are for the purpose of Victim/Trafficker Identification, and they typically employ open-source intelligence (OSINT) methods to analyse the data (OSCE, 2020). OSINT encompasses a wide range of publicly available sources, such as social media posts and escort advertisements (Campana, 2022). This thesis focusses on Victim/Trafficker Identification tools, as they are the most commonly used tools and align with the research objectives of identifying recruitment activities online.

Victim/Trafficker Identification tools for detecting sex trafficking primarily analyse escort and classified advertisements to distinguish those promoting minors

from those involving adults. Determining whether an adult is being trafficked presents significant challenges, in contrast to cases involving children, where any minor engaged in commercial sexual activity is considered a victim. This distinction may explain why most of the solutions focus on identifying children.

Table 2.1 presents examples of Victim/Trafficker Identification tools, most of which are considered "reactive", as they are employed after victims have already been recruited.

**Table 2.1:** A Summary of Anti-trafficking Software

| Name | Description |
|---|---|
| PhotoDNA | Microsoft developed this tool to identify victims of human trafficking and child sexual exploitation content online. It finds copies of child sexual exploitation images online by comparing them against other images using unique digital signatures. Once the images match, they report to the National Centre of Missing and Exploited Children (OSCE, 2020). |
| Tellfinder | Formerly known as Memex, incorporates components that assist law enforcement in analysing information to help identify cases of trafficking online. It has a specific program called DIG that collects information from various sources on online platforms beyond what standard search engines can collect, such as the deep web. This information helps to develop patterns and connections to identify suspected trafficking cases and organisations or critical people involved in trafficking (OSCE, 2020; Pendergrass, 2018). Tellfinder can also be used for academic research, policy development, and intelligence gathering by other stakeholders (TellFinder, n.d.). |
| Spotline | Thorn developed this tool to help law enforcement identify victims and traffickers of child sex trafficking online. It collects and analyses information on the Internet, focussing on escort and adult websites to identify trafficking advertisements. The tool helps identify six children daily and saves 60% of investigation time (Thorn, 2018; Pendergrass, 2018). |
| ePoolice | Assists law enforcement with their operations by identifying and predicting organised crime, including sex trafficking. It gathers data from OSINT, police sources, and public and government databases (Centric, 2022; Brewster et al., 2014). |
| TrafficBot | The tool generates useful information that law enforcement uses to identify and combat the sex trafficking of children online. It extracts information from six classified websites and two John Boards (websites used by clients to exchange information on children) (H. Wang et al., 2012). |

Continued on next page

**Table 2.1 – continued from previous page**

| Name | Description |
|---|---|
| idTraffickers | A tool developed for law enforcement to help them to sort through online data to identify traffickers and victims. The tools help the officers analyse large quantities of online sex ads. (OSCE, 2020; idTraffickers, 2022). |
| STOP | A tool by NORC at the University of Chicago analyses information from escort ads and websites to help law enforcement with their investigations and attempts to end sex trafficking. It identifies patterns from the data and produces reports for officers (NORC, n.d.). |
| Traffic Jam | Law enforcement uses the tool to find patterns that can assist them in identifying and stopping human trafficking. It enables them to identify potential victims and find clues about their whereabouts, which can lead to their rescue. In addition, it helps collect information that assists in building cases against traffickers (Pendergrass, 2018; Freedom Signal, n.d.). |
| TraffickCam | This tool creates a database of hotel images to help law enforcement identify the location of victims during investigations. It works by having people upload images of their hotels to match them with images of victims in hotels (OSCE, 2020; TraffickCam, 2022). |
| Freedom Signal | A tool that identifies victims of sex trafficking by analysing the Internet platforms. It lets different parties communicate with the victims once the tool identifies them (Freedom Signal, n.d.). |

Beyond the growing need to combat online exploitation, the predominance of Victim/Trafficker Identification tools focused on the exploitation stage stems from a deep understanding of how traffickers exploit victims and how anti-trafficking stakeholders detect criminal activities at this stage. This knowledge has facilitated the development of anti-trafficking tools that support stakeholders in their efforts. However, as noted in Section 2.3.2, the same depth of understanding does not exist for online recruitment. Therefore, alongside deepening knowledge of recruitment, understanding how stakeholders investigate recruitment can contribute to the development of tools that enable anti-trafficking workers to identify recruitment activities more effectively.

Various stakeholders use Victim/Trafficker Identification tools to assist with their work. These stakeholders include anti-trafficking organisations, law enforcement agencies, and individuals who have been affected by trafficking. Law enforcement

agencies are the main users of Victim/Trafficker Identification tools and employ them to support their investigations. Anti-trafficking organisations also use Victim / Trafficker Identification tools to facilitate victim rescue and report cases (OSCE, 2020).

Similarly, multiple stakeholders, including the private sector, non-governmental organisations, government entities, and academic institutions, develop Victim/Identification tools. However, there are concerns about the motivations underlying the development of these tools, particularly when initiated by individuals or private entities (Musto et al., 2020). Although the motivations behind the development of certain private sector tools may not always be transparent, these tools have exerted a significant influence on the operations of law enforcement agencies in both the US and the United Kingdom (UK) (Pendergrass, 2018).

The following section discusses the impacts of these tools.

## 2.5 The Impact of Anti-Trafficking Tools

Anti-trafficking tools automate data crawling and analysis, substantially reducing processing time and minimising the likelihood of errors. These tools are crucial to efficiently summarising potential trafficking cases for further investigation. In particular, automating image analysis has proven to be highly advantageous in terms of time efficiency, as manual examination is often prone to errors and biases (Pendergrass, 2018). The tools highlighted in Table 2.1 such as Spotlight, PhotoDNA, and TrafficJam have each contributed meaningfully to the fight against human trafficking. Spotlight has helped identify over 17,000 children (Spotlight, n.d.), while PhotoDNA has enabled Google to remove 270,000 users involved in CSAM (Flores, 2022). In 2020 alone, TrafficJam is estimated to have saved more than 70,000 hours of investigative work (Marinus Analytics, 2022).

Although anti-trafficking tools provide significant value, they also have various limitations. One of the main challenges is obtaining accurate data for these tools. Previous research (Musto, 2020) has stated that most of the data in this field are unverified and unreliable, which can lead to false positives. Another limitation

is the focus on specific regions, allowing traffickers to exploit an unmonitored geographical area.

Multiple studies have also raised concerns about possible privacy issues associated with their use (Latonero and Movius, 2011; Latonero, 2012; Deeb-Swihart et al., 2019; Pendergrass, 2018; Ediae and Kuteesa, 2024; Rivas et al., 2024). The integration of sex worker data during data collection, processing, and storage raises substantial privacy concerns (Deeb-Swihart et al., 2019). This issue is particularly complex in countries where sex work is illegal, such as the US, as integrating data from voluntary sex workers may have severe negative consequences (Musto, 2020). Privacy concerns are especially pressing in this context because mishandling data could lead to unintended harm, including public exposure of victims' information.

Several studies emphasise the importance of incorporating privacy considerations into anti-trafficking tools and investigations to mitigate these risks (OSCE, 2020; Latonero, 2012; Deeb-Swihart et al., 2019; Pendergrass, 2018). However, the balance between privacy and security remains a critical challenge, as measures such as encryption can also be exploited by traffickers to evade detection (Anderson et al., 2024). Furthermore, regulations designed to protect privacy can sometimes obstruct efforts to identify victims and traffickers through such tools, making law enforcement interventions more complex (Campana, 2022).

The use of technology in anti-trafficking efforts also introduces other risks beyond privacy, which can have broader negative implications for victims and society. Some studies stress the need to better understand these risks and develop measures that guide the responsible development of these tools (OSCE, 2020; Latonero and Movius, 2011; Latonero, 2012; Deeb-Swihart et al., 2022). However, there is a limited body of research that thoroughly examines these concerns (Latonero, 2012). Previous research (Deeb-Swihart et al., 2019) investigated the technological requirements of law enforcement in the fight against human trafficking and identified various challenges that influence tool design and use. However, this study only briefly addressed potential unintended consequences and did not thoroughly examine them.

To effectively mitigate the risks posed by these tools, it is essential to identify potential harm to victims and society, ensuring that appropriate measures are established (OSCE, 2020; Latonero, 2012; Taylor and Latonero, 2018). Whether developers of anti-trafficking tools actively incorporate such safeguards during the development process remains uncertain. Researchers such as Taylor and Latonero emphasise the need for greater awareness of the risks and harms associated with these technologies and advocate for policies and procedures that safeguard fundamental human rights within the anti-trafficking domain (Taylor and Latonero, 2018). Data mishandling can have severe consequences, including public exposure of victims' identities, exacerbating the risks they already face (Deeb-Swihart et al., 2022). Addressing these concerns requires continuous research, policy discussions, and collaboration between stakeholders to ensure that developers take ethical considerations into account when designing these tools. Establishing standardised measures for data protection, ethics and informed consent is essential to align anti-trafficking technologies with ethical and legal frameworks (OSCE, 2020).

## 2.6 Related Work on Technological Solutions to Detect Online Trafficking

Over the years, researchers in academia have leveraged various technologies to develop solutions that assist law enforcement and other stakeholders in analysing large volumes of online data to identify trafficking activities on the Internet. They have also compared different technological approaches to determine which techniques yield the highest accuracy in detecting trafficking activity.

This section reviews academic articles on the detection of human trafficking activity and examines the technologies that have been employed. It also highlights the data sources and acquisition methods used for trafficking detection. These articles were identified using the approach described in the Introduction section of this chapter.

### 2.6.1 Data Sources and Acquisition

Detecting and analysing online human trafficking through computational approaches, especially using artificial intelligence, requires vast amounts of data. Concerns have been raised regarding the limited availability of data in this field, which impacts the scope of the work that can be conducted and the methodologies that can be employed (Dorton, 2019; Ediae and Kuteesa, 2024; Rivas et al., 2024; Amin, 2010). Nevertheless, several datasets have been developed. Most academic studies in this field use data from escort websites as their main source (Dubrawski et al., 2015; Alvari et al., 2017; Tong et al., 2017; Alvari et al., 2016; Zhu et al., 2019; Wang et al., 2020; Paula et al., 2023; Esfahani et al., 2020; J. Liu et al., 2023; Summers et al., 2023; Wiriyakun and Kurutach, 2021; Vajiac et al., 2023; Nair et al., 2022; Ardakani, 2020; Freeman et al., 2022). For example, a previous study (Tong et al., 2017) developed a dataset called the Trafficking-10K dataset, comprising 10,000 annotated escort advertisements from Backpage. This dataset included advertisements with images and text from the US and Canada and was annotated by two experts with more than five years of experience and one expert with at least one year of experience (Tong et al., 2017). Esfahani et al. labelled advertisements on Backpage and Eroticmugshots by cross-referencing phone numbers extracted from advertisements with a database of phone numbers associated with trafficking victims (Esfahani et al., 2020). If a number in an advertisement appeared in both sets, the ad was classified as trafficking-related.

Several studies have used social media data as a source (Hernandez-Alvarez, 2019; Simonson, 2021). Simonson developed a labelled dataset of Instagram posts employing a semi-supervised methodology (Simonson, 2021). A semi-supervised approach involves training an algorithm on both labelled and unlabelled data. It falls between supervised and unsupervised learning. In collaboration with an expert, the researcher manually labelled approximately 700 posts with the objective of automatically labelling 50,000 posts in total (Simonson, 2021). The FastText CBOW Model with six clusters demonstrated the highest cross-validation accuracy during the initial label propagation, and this approach was subsequently employed

to label 12,578 posts. However, due to time constraints, the researcher did not complete the labelling process for the remaining 37,194 posts.

In addition to the datasets created for sex trafficking research, Tobey developed a dataset on Illicit Massage Businesses based on YELP reviews (Mayorga and Özaltın, 2023). The researchers used active learning to label the data. Active learning enables a model to learn from a limited amount of labelled data and to complete the labelling process (Mucinno, 2019; Jacobs et al., 2021). Active learning allows a model to learn from a small set of labelled data and finalise the labelling (Mucinno, 2019; Jacobs et al., 2021). This strategy, which started with manual labelling, resulted in a labelled dataset comprising 1,735 reviews.

One of the necessary steps in creating anti-trafficking datasets is data annotation. Data annotation is the process of manually labelling data in the form of text, images, or videos (TELUS International, 2021. Most research involves text annotation because studies predominantly focus on textual information rather than images. Text annotation is the process of adding metadata to the data to enhance their semantic values. This process ensured that the prediction of the model was accurate and produced reliable results. Due to the large volume of data, text annotation can be labour intensive, necessitating human experts to label the data. Several studies (Alvari et al., 2017; Simonson, 2021; Tong et al., 2017) used semi-supervised learning for text annotation to address these challenges and develop data for anti-trafficking models. Semi-supervised learning for text annotation involves training an algorithm on labelled and unlabelled data, which falls between supervised and unsupervised learning (Arora, 2019; Dickson, 2021). This approach enables human experts to label only a subset of the data points, after which the model labels the remaining data.

The Labelled data are not guaranteed to be accurate or reliable. To mitigate the number of erroneous data points, additional data can be labelled (Mucinno, 2019); however, this process is time consuming. Inter-Annotator Agreement (IAA) has been used in some studies to assist in validating their expert annotations (Tong et al., 2017). IAA quantifies the agreement between multiple annotators when categorising data into specific categories (Bruijn, 2020). In instances where there are at least two

annotators, Cohen's Kappa can be employed to calculate the agreement reliability. This method is applicable to categorical items and facilitates the assessment of classification models (Bruijn, 2020; Widmann, 2020). An alternative approach is the use of Fleiss' Kappa, which determines the reliability of the agreement for a specified number of annotators (Bruijn, 2020).

### 2.6.2 Detecting Interned-based Trafficking using Machine Learning and Natural Language Processing

Current research in this field employs various approaches to identify trafficking activities. Several studies (Dubrawski et al., 2015; Lugo-Graulich et al., 2024; Alvari et al., 2017; Giommoni and Ikwu, 2024; Zhou et al., 2024; Hultgren et al., 2016; Whitney, 2017; Hernandez-Alvarez, 2019; Poelmans et al., 2010) have used keywords, or combinations thereof, to detect trafficking activities with machine learning models. Traffickers often employ specific keywords, emojis, and other indicators in advertising trafficking (United Nations Office on Drugs and Crime, 2018; Lugo-Graulich et al., 2024; Giommoni and Ikwu, 2024; Zhou et al., 2024; Hernandez-Alvarez, 2019). Previous research, in conjunction with law enforcement and anti-trafficking organisations, has contributed to the development of a list of keywords and emojis that developers use to identify trafficking. Keywords can consist of a single word or phrase. However, traffickers continuously adapt to avoid detection by altering keywords and employing complex obfuscation techniques, preventing ML models from identifying advertisements (Tong et al., 2017). Several studies have been conducted to counteract these evasion techniques. Lugo-Graulich (Lugo-Graulich et al., 2024) presented a novel approach that corrects misspelled words in advertisements by replacing them with terms that capture their intended meanings. This method also accounts for emojis. There are warnings against relying solely on prescribed indicators to detect online advertisement for human trafficking (Campana, 2022). It is recommended to incorporate additional data sources to help determine whether advertisements are linked to human trafficking. However, it is challenging to identify multiple data sources in this field. Therefore, researchers should explore

alternative methods for detecting trafficking activities that do not rely solely on indicators and that can automatically adapt to traffickers' obfuscation techniques.

Several studies (Ramchandani et al., 2021; Dubrawski et al., 2015; Simonson, 2021; Tong et al., 2017; Alvari et al., 2016; Zhu et al., 2019; Wang et al., 2020; Summers et al., 2023) **have used NLP to identify trafficking activities**. Dubrawski et al. used keywords, NLP feature sets, and Regular Expression (RegExp) to train multiple Random Forest models to identify sex trafficking advertisements on the Internet (Dubrawski et al., 2015). NLP feature sets demonstrated greater accuracy in identifying potential sex trafficking advertisements than keywords and regular expression approaches. The keyword-based model exhibited superior performance compared to RegExp, although the difference in performance was minimal. Using the NLP approach, they implemented an unsupervised bag-of-words to augment the features in the data, followed by Principal Component Analysis for dimensionality reduction. The advantages of employing NLP feature sets include their independence from human expertise in contrast to the use of keywords or RegExp (Dubrawski et al., 2015).

Alvari et al. employed a non-parametric learning approach to identify sex trafficking advertisements on escort websites (Alvari et al., 2016). They used an unsupervised filtering approach to identify data likely pertaining to sex trafficking, and subsequently trained a semi-supervised learner on data labelled by experts to classify unlabelled data. The unsupervised learning approach involves the application of Latent Dirichlet Allocation (LDA) to obtain representative topics that facilitate the distinction between non-trafficking content and potential trafficking content. LDA is a topic modelling method that identifies topics from data, which helps to identify hidden patterns (Barber, n.d.). Human experts labelled a portion of the filtered data, which were subsequently input into a semi-supervised learner to classify the unlabelled advertisements.

Several studies (Paula et al., 2023; Esfahani et al., 2020; J. Liu et al., 2023) **have used state-of-the-art NLP approaches, such as pre-trained neural language models.** Esfahani et al. developed a collection of features comprising those

obtained using the LDA model, FastText, and Bidirectional Encoder Representations from Transformers (BERT) (Esfahani et al., 2020). This set of collective features demonstrated superior performance compared with feature sets from previous studies (Esfahani et al., 2020). The collective feature set was used in the predictive models to identify trafficking activities. Liu et al also used recent state-of-the-art approaches such as RoBERTa and DeBERTa to extract meaningful insights from noisy data containing sensitive and private information, which facilitated the detection of sex trafficking advertisements (J. Liu et al., 2023). The authors used two human trafficking datasets and two public datasets, CoNLL2003 and WNUT2017, to train models that extract person names from escort advertisements. There is still a need to explore more advanced models, such as **LLMs**, to evaluate their performance in detecting human trafficking. Additionally, more work is needed to assess pre-trained models on existing popular datasets, as few studies have utilised them despite their excellent performance in other fields.

Although most studies focus on text, **several pieces of work have incorporated multimodal data** (Whitney, 2017; Whitney et al., 2018; Tong et al., 2017; Summers et al., 2023.) One such example is the work of Tong et al., who developed a deep multimodal model called Human Trafficking Deep Network (HTDN) to detect advertisements on escort websites (Tong et al., 2017). The authors created the Trafficking-10K dataset described above and used it to train the HDTN. The HDTN comprises two components: the language and visual components, which process text and images, respectively. The model initially generates word embeddings using the Skip-gram model. Skip-gram models learn word representations by predicting context words given a target word. HDTN subsequently employs Long Short-Term Memory (LSTM) as the deep learning architecture for the language network. LSTM possesses a cell memory that allows it to capture long-term dependencies (Madhavan, 2020), allowing HDTN to retain relevant information when encountering irrelevant details in advertisements. Similarly to the language network, the visual network uses the T-VGG network to generate visual representations. T-VGG is a Convolutional Neural Network architecture that is fine-tuned from the VGG

network. Language and visual networks are fed into a convolutional decision network that classifies whether an advertisement is related to trafficking. HTDN demonstrated better performance compared to models that only used text or images when evaluated against previous baselines.

Zhu et al. developed an approach that exceeded the performance of the HTDN system despite using only textual data for training (Zhu et al., 2019). Their method involved training a language model using text in advertisements and incorporating a phrase detection algorithm to identify phrases commonly used in trafficking advertisements. Subsequently, they implemented an N-gram language model (KenLM) specifically tailored to adult service advertisements. Following the application of a trafficking dataset, they employed a feature selection method using an L1 linear SVM on TF-IDF vectors to remove irrelevant features. The model ultimately employs logistic regression to classify advertisements as high- or low-risk, ranking them based on their probability of being trafficking-related.

Wang et al. developed an ordinal regression neural network using advanced deep learning techniques to identify sex trafficking advertisements (Wang et al., 2020). The researchers compared their approach with the HDTN system described above and demonstrated that their approach exceeded the HTDN in terms of classification accuracy. Their model comprised three components: an input layer that generated word embeddings using a Skip-gram model, a deep learning architecture that extracted summary features from sentences using a Gated-Feedback Recurrent Neural Network (GF-RNN), and an output layer customised with multi-label logistic regression for ordinal regression. This study used an approach similar to that of Tong et al., which learns word embeddings from escort advertisements (Tong et al., 2017). The GF-RNN used by Wang et al. is a refined version of the Long Short-Term Memory (LSTM) model (Wang et al., 2020) employed by Tong et al. in the HTDN (Tong et al., 2017). GF-RNN enhances the ability of a recurrent layer to capture dependencies at various time scales, enabling the model to automatically adjust its structure based on the input (Chung et al., 2015). A few studies demonstrated that

models that use only textual data still have the ability to produce superior results compared to models that use multimodal data (Zhu et al., 2019; Wang et al., 2020).

Wiriyakun et al. propose a feature selection approach using Local Interpretable Model-Agnostic Explanations (LIME) to enhance the performance of machine learning models in detecting suspicious online advertisements (Wiriyakun and Kurutach, 2021). Their method identifies important words for training classification models, leading to improved accuracy. When applied to the Trafficking-10k dataset, it outperformed previous techniques, particularly in Naïve Bayes, Decision Tree, and Neural Network models.

**Identifying traffickers has been another main focus,** with several studies using mathematical frameworks to analyse law enforcement data to detect suspected traffickers. Poelmans et al. integrated Formal Concept Analysis (FCA) with expert input to detect human trafficking activities from unstructured data (Poelmans et al., 2010). This approach analysed 266,157 reports, enabling law enforcement to visualise and analyse complex information, which subsequently led to investigations and the arrest of traffickers. Poels et al. built on this previous work, employing Temporal Component Analysis (TCA) to identify suspected traffickers from police reports and analyse their behaviours and social networks over time (Poelmans et al., 2010). TCA is advantageous because it incorporates the temporal component, allowing concepts to be analysed longitudinally, unlike FCA. This enables stakeholders to better comprehend how traffickers' methods evolve, which can aid in the development of more effective strategies. These strategies emphasise the importance of analysing existing work from anti-trafficking stakeholders, as they can provide essential information about human traffickers and victims. Existing work, such as reports, offers an additional source of data.

In addition, several studies (Arenas et al., 2015; Vajiac et al., 2022, 2023; Nair et al., 2022; Ardakani, 2020) **have focused on identifying organised trafficking networks and online behaviour**. Many of these studies (Arenas et al., 2015; Vajiac et al., 2022, 2023; Nair et al., 2022) have used social network analysis and clustering approaches, such as multi-level clustering, duplicate information

clusters, and joint clustering, to visualise online escort advertisements and social media data to detect sex trafficking and organised communities. For example, Freeman et al. developed a tool that employed various techniques, such as network analysis and predictive modelling, to link data from multiple online platforms and identify organised sex trafficking networks (Freeman et al., 2022). This tool enabled law enforcement to filter data by time and location, which proved crucial in identifying trafficking activities. This contributed to more than 40 arrests. Vajiac presented a scalable, effective, parameter-free, and interpretable method for detecting micro-clusters of near duplicate documents, primarily focusing on detecting human trafficking in escort advertisements and Twitter bot activity (Vajiac et al., 2022). The paper emphasises that identifying clusters of related documents is essential for human trafficking detection, as similar escort ads often indicate organised trafficking activity.

The studies discussed above have focused predominantly on advertisements on escort websites. However, several studies (Hernandez-Alvarez, 2019; Simonson, 2021; Arenas et al., 2015; Burbano and Hernandez-Alvarez, 2018) **have also used data from social media platforms.** Burbano and Hernandez-Alvarez developed a Spanish corpus based on social networking websites and newspapers and subsequently employed ML algorithms to create a predictive model that identifies human trafficking patterns (Burbano and Hernandez-Alvarez, 2018). The authors established an approach to identify the most efficient features to train ML algorithms to identify human trafficking patterns in the data. Simonson similarly employed a semi-supervised approach to identify Instagram posts that indicate belonging to the commercial sex industry (Simonson, 2021). The author used the FastText and Word2vec models to embed Instagram posts in 100-dimensional vectors. FastText demonstrated superior performance over Word2vec, achieving 98% accuracy compared to Word2vec 50%. Simonson subsequently applied the K-means model to group the data and used cross-validation to verify the accuracy of the labels (Simonson, 2021). This research aims solely at identifying posts in the commercial sex industry due to the difficulty in determining whether a post is related to trafficking.

A few studies have also aimed to **investigate and identify trafficking during the recruitment stage** (Ramchandani et al., 2021; Zhou et al., 2024; Volodko et al., 2020). Ramchandani employed NLP, active learning, network analysis, and domain expertise to detect recruitment patterns indicative of human trafficking within the commercial sex supply chain (Ramchandani et al., 2021). The study analysed job and commercial sex advertisements from websites primarily based in the US, Canada, the UK, and Australia. These advertisements were linked to potential trafficking activities by comparing shared metadata, such as phone numbers, emails, usernames, and social media handles, between job advertisements and commercial sex advertisements. In addition, social network analysis was used to map geographical trafficking routes, providing information on where victims were recruited and subsequently advertised for exploitation.

A 2024 study by Zhou analysed a quarter million job advertisements in the US targeting Chinese-speaking immigrants to identify potential human trafficking activities (Zhou et al., 2024). The study collected and categorised data while simultaneously conducting a linguistic analysis of keywords, job descriptions, temporal patterns, and contact information associated with suspicious advertisements. Lexicons based on trafficking research and hypothetical victim scenarios, including terms such as "visa sponsorship" and "under the table", were employed to identify advertisements that may appeal to vulnerable populations. Furthermore, phone numbers were a significant indicator of suspicious job ads, as they were commonly used in high-risk listings and linked to potential trafficking activities. This study discussed the role of job advertisements in facilitating human trafficking and highlighted the need for further research and targeted interventions to address this issue (Zhou et al., 2024).

**There is a lack of research on the use of computational methods to detect fraudulent job opportunities** compared to exploitation, which limits the understanding of optimal approaches. The Trafficking-10K dataset developed by Tong et al. has demonstrated significant influence in the field (Tong et al., 2017), with a lot academic studies (Zhu et al., 2019; Wang et al., 2020; Summers et al., 2023; Wiriyakun and Kurutach, 2021; Vajiac et al., 2023) using this dataset to

identify approaches that enhance the accuracy of sex trafficking detection during the exploitation stage, thereby contributing to the development of more effective anti-trafficking software. It is imperative that similar labelled datasets for fraudulent jobs are developed and subjected to experimentation.

In addition to the work on sexual exploitation, several studies (Mayorga and Özaltın, 2023; Li et al., 2023; Diaz and Panangadan, 2020a) **have also focused on identifying illicit massage parlours.** Most of these studies analyse online business and recommendation websites, such as Yelp, to identify establishments that may be associated with human trafficking. These investigations (Mayorga and Özaltın, 2023; Li et al., 2023; Mucinno, 2019; Diaz and Panangadan, 2020a, 2020b) generated datasets based on Yelp data and employed traditional approaches such as lexicon-based models, Bag-of-Words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF), Doc2Vec, and Word2Vec, as well as more advanced models such as BERT. Ruoting et al. examined customer reviews on public websites such as Yelp, using a custom lexicon of trafficking-related terms in conjunction with embeddings from the BERT and Doc2Vec models (Li et al., 2023). Lexical-based models demonstrated high precision, while embedding models exhibited higher recall. There are opportunities for further research in this domain, as techniques employed to identify online sex trafficking may also be applicable to this field.

## 2.7 Anti-trafficking Policies and Online Legislation

In response to the threats present on online platforms, various legislative measures have been implemented to address the misuse of these platforms and protect individuals from criminal activities, such as human trafficking. Most legislation specifically targeting human trafficking tends to focus on sex trafficking, as it represents the most prevalent form of human trafficking conducted using online channels.

The Communications Decency Act of 1996 conferred immunity to Internet Service Providers and online platforms regarding user-generated content, subsequently eliciting criticism for facilitating human trafficking advertisements online (Cotterill,

2023). In response, the US introduced the SAVE Act in 2015, which imposed responsibility on platforms for trafficking content; however, it had significant limitations. Online platforms were initially accountable for the content created by users; however, this accountability was lifted if the platforms made any effort to moderate the content, thus creating possible legal loopholes. To address these limitations, the FOSTA-SESTA package was enacted in 2018, which amended the CDA to require that online platforms be legally accountable for all user-generated sex trafficking content (Cotterill, 2023; Flynn, 2021; Romano, 2018).

Similarly, in Europe, the Digital Services Act mandates that online platforms in the European Union (EU) address illegal content, including human trafficking. This is done not by monitoring user content, but by removing flagged content and using transparent automated moderation tools (European Commission, n.d.; Pirkova, 2022). The UK Online Safety Bill also requires platforms to ensure user safety by proactively monitoring and removing illegal content while considering user privacy (GOV.UK, 2023). Similarly, the Australia Online Safety Act of 2021 requires that platforms protect users from illegal content (ActiveFence, 2024; OneTrust Data Guidance, 2022). All three regulations impose substantial penalties for non-compliance.

In Africa, Kenya implemented the Computer Misuse and Cybercrimes Act No. 5 of 2018 (the Act) to facilitate the prevention, prosecution, detection, investigation, and punishment of cybercrime (National Computer and Cybercrime Coordination Committee, n.d.). The Act also mandates the protection of accountability, integrity, and availability of computer systems, programmes, and data. South Africa introduced the Films and Publications Act in 1996 and amended it in 2022 (Gazette and Notice, 2022). This legislation aims to ensure online safety by mitigating online harm, including provisions that require online distributors and ISPs to protect the general public, particularly children, from explicit and harmful content.

In 2024, the African Union Ministry introduced the world's first policy framework for implementing children's rights in the digital space: the Child Online Safety and Empowerment Policy (African Union, 2024; 5Rights Foundation, 2023; NETalent,

2024). This policy aims to protect children from harm online and offline, which has become increasingly prevalent. This policy assesses the benefits and potential risks of digital platforms for children in Africa and identifies key issues and policy objectives (African Union, 2024; 5Rights Foundation, 2023; NETalent, 2024). Furthermore, this policy seeks to assist African countries in implementing legislation aimed at protecting 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 (The UN Agency for Digital Technologies, 2021). Only a limited number of countries have enacted laws that address online harm such as cyberbullying and human trafficking. As countries in Africa strive to establish legislation for digital platforms, policymakers would benefit from access to more comprehensive information on the role of the Internet in facilitating human trafficking in Africa to develop the necessary measures to address this issue.

## 2.8 Summary and Current Limitations

The preceding literature review has examined human trafficking, technology-facilitated trafficking, anti-trafficking tools and their efficacy, current computational approaches to identify human trafficking, and existing legislation that helps combat this crime. This review of the literature has identified gaps in the literature on online trafficking. The subsequent sections discuss the four primary limitations that will be addressed in this thesis.

### 2.8.1 Limited Studies that Assess the Impact of Anti-Trafficking Tools

As observed, a wide range of stakeholders, including law enforcement agencies, governments, and non-governmental organisations, use software tools to detect human trafficking. These software tools are primarily focused on identifying victims and traffickers and raising awareness (OSCE, 2020). Such tools offer several advantages, such as time efficiency and enhanced work process optimisation.

However, there are concerns about their use, particularly in relation to data privacy, bias, and potential risks (Latonero, 2012; Ediae and Kuteesa, 2024; Deeb-Swihart et al., 2022). The issues of privacy and bias can affect victims in negative ways. Unfortunately, only a few studies have thoroughly investigated these concerns (Latonero and Movius, 2011). It is crucial to explore and address the potential risks and negative effects that these tools can have on victims and society (OSCE, 2020; Latonero and Movius, 2011; Musto et al., 2020).

In addition, there is a notable absence of measures to guide the development of anti-trafficking software. A lack of such measures can place victims and other stakeholders at risk (OSCE, 2020; Ediae and Kuteesa, 2024). Identifying the risks and challenges involved in developing anti-trafficking software is crucial to devise measures to mitigate them during and after the design phase and to implement solutions that consider various vulnerable individuals in society (OSCE, 2020).

Although it is imperative to monitor and identify trafficking activities, it is equally crucial to ensure that the tools developed do not adversely affect victims and the general public. Developers should be aware of the potential risks associated with developing such tools and incorporate the necessary measures to mitigate any identified risks.

### 2.8.2 Limited Work on Online Recruitment

Traffickers commonly use online platforms to recruit and advertise sex trafficking victims, then subsequently connect them with buyers seeking services, such as prostitution (OSCE, 2020; Latonero and Movius, 2011; Department of State, 2023; Ramchandani et al., 2021; United Nations Office on Drugs and Crime, 2018; Europol, 2014; Zhou et al., 2024; US Department of State, 2019). Several studies have examined how traffickers exploit victims on online platforms (Latonero and Movius, 2011; Mitchell and Boyd, 2014; Hultgren et al., 2018; Dubrawski et al., 2015; Mayorga and Özaltın, 2023; Alvari et al., 2017; Giommoni and Ikwu, 2024; Leary, 2014; Hultgren et al., 2016; Hernandez-Alvarez, 2019; Simonson, 2021; Sarkar, 2015; Arenas et al., 2015; Nair et al., 2022). As seen, these studies revealed

the patterns, keywords, and indicators of sex trafficking advertisements on these platforms. Only a few studies have discussed and described online recruitment methods such as relevant keywords and indicators (Anthony, 2018; Ramchandani et al., 2021; Zhou et al., 2024; Volodko et al., 2020), although it has been one of the most widely used recruitment methods in recent years (Progetto Tenda et al., 2021; Feehs and Wheeler, 2020; United Nations Office on Drugs and Crime, 2020; Zhou et al., 2024; Gezinski and Gonzalez-Pons, 2022).

Online recruitment facilitates the expansion of criminal enterprises by allowing traffickers to target a significantly larger population than is possible by conventional means (Europol, 2014). A lack of understanding of this issue leads to inefficient solutions to address this problem. Additional studies are necessary to analyse and understand online recruitment, particularly regarding how traffickers target and recruit victims on social media platforms (Department of State, 2023; Anthony, 2018; Mitchell and Boyd, 2014; Thorn, 2018; Greiman and Bain, 2013; Ibanez and Suthers, 2014; Latonero, 2012; Zhou et al., 2024). The 2023 US Department of State report urged governments and technology companies to recognise the recruitment of vulnerable individuals online and collaborate on solutions to prevent these recruitment practices (Department of State, 2023).

### 2.8.3 Lack of Anti-trafficking Technologies that Identify Online Recruitment

The majority of technologies in this field focus on detecting human trafficking during the exploitation stage (Dorton, 2019; Latonero and Movius, 2011; Anthony, 2018; Dubrawski et al., 2015; Mayorga and Özaltın, 2023; Thorn, 2018; Latonero, 2012; Leary, 2014; Hultgren et al., 2016; Hernandez-Alvarez, 2019; Simonson, 2021; Sarkar, 2015; Arenas et al., 2015; Whitney et al., 2018; Tong et al., 2017; H. Wang et al., 2012; Silva et al., 2014; Alvari et al., 2016; Zhu et al., 2019; Wang et al., 2020; Paula et al., 2023; Esfahani et al., 2020; J. Liu et al., 2023; Summers et al., 2023; Wiriyakun and Kurutach, 2021; Vajiac et al., 2022, 2023; Nair et al., 2022; Ardakani, 2020; Li et al., 2023; Freedom Signal, n.d.; Burbano and

Hernandez-Alvarez, 2018; Diaz and Panangadan, 2020a). Few of these technologies are designed to identify online recruitment activities (Ramchandani et al., 2021; Zhou et al., 2024; Volodko et al., 2020).

Recommendations have been made to develop a tool that identifies fraudulent job postings to address online recruitment (Department of State, 2023; Zhou et al., 2024; Simonson, 2021). The creation of such a tool is crucial, because fraudulent job postings are a key tactic in online recruitment (Feehs and Wheeler, 2020. By detecting and understanding online recruitment practices, victims can be prevented from experiencing traumatic situations (Volodko et al., 2020). Previous researchers stated that

> "If it were possible to identify potential trafficking more effectively at the recruitment stage, there would be considerable potential for prevention, detection, disruption and early intervention" (Volodko et al., 2020).

The lack of data is one of the primary factors that has restricted progress in this particular domain (Dorton, 2019; Ediae and Kuteesa, 2024; Rivas et al., 2024; Amin, 2010). This scarcity of data hinders research on online recruitment and interferes with the development of technological solutions to address online recruitment. More publicly labelled data sets are needed to develop automated solutions to detect this crime (Dorton, 2019).

### 2.8.4 Limited Use of State-of-the-art Methods in Natural Language Processing

Studies in this field apply various NLP techniques to detect trafficking activities. Some of these studies have used the 'bag of words' and TF-IDF methodologies (Dubrawski et al., 2015; Alvari et al., 2016; Zhu et al., 2019); however, these methods do not consider the relationships between words or adequately consider their context. To address this limitation, several studies have generated word embeddings using Word2vec, skip-gram, and CBOW models (Simonson, 2021; Tong et al., 2017; Wang et al., 2020). However, in recent years, new transformer-based models have been introduced, such as BERT and RoBERTa. These models achieve state-of-the-art

performance in various NLP tasks, such as text classification (Islam et al., 2024; Vaswani et al., 2017; Devlin et al., 2018).

Most studies have not used state-of-the-art methods, with the exception of a limited number that incorporate these methods (Paula et al., 2023; Esfahani et al., 2020; J. Liu et al., 2023). Furthermore, large language models have also been introduced, such as GPT models, and, to the best of my knowledge, these models have not been leveraged or experimented with in the technology-facilitated trafficking field. These advanced techniques have the potential to enhance content classification based on various contexts of words, thereby improving performance and precision. Considering that traffickers generate content that misleads potential victims, law enforcement, and automated systems, it is necessary to employ methods that recognise the various contexts of words used by traffickers for accurate content classification. Moreover, most state-of-the-art methods have not been implemented in the online recruitment phase, which presents the opportunity to investigate their efficacy in addressing this issue.

## 2.9 Direction of Thesis

Based on the literature review and the limitations identified, this thesis formulated the research questions discussed in Section 1 to address the aforementioned limitations. Furthermore, the following decisions were made regarding the direction of this thesis:

1. The decision to focus on sex trafficking was based on three main reasons: 1) it is the most frequently detected type of human trafficking, 2) women and girls, who constitute the majority of trafficking victims, are predominantly exploited through this form, and 3) sex trafficking is the type of trafficking most significantly affected by online technology (Ramchandani et al., 2021; United Nations Office on Drugs and Crime, 2020, 2018; Mitchell and Boyd, 2014; United Nations: Office on Drugs and Crime, 2023; United Nations Office on Drugs and Crime, 2012).

2. The US was initially selected as the geographic focus due to the prevalence of online recruitment and the availability of data sets in the region (Feehs and Wheeler, 2020; Ramchandani et al., 2021; Thorn, 2018). However, this decision was subsequently changed following the initial phase of the thesis based on the findings and discussions with the participants. This modification is explained in more detail in Chapter 3.

# 3

# Assessing Anti-Trafficking Software

## Contents


**3.1 Introduction . . . . . 58**
**3.2 Methodology . . . . . 59**
3.2.1 Responsible Research and Innovation . . . . . 60
3.2.2 Anticipate: Analysing the Consequences of Anti-trafficking Tools . . . . . 61
3.2.3 Reflection: Examination of Motivations and Benefits . . 62
3.2.4 Engagement: Involving Diverse Stakeholders . . . . . 63
3.2.5 Act: Integrating Stakeholder Perspectives . . . . . 67
3.2.6 Comparative Approach: US and UK Contexts . . . . . 67
**3.3 Research Findings . . . . . 67**
3.3.1 Challenges Associated with Developing Anti-Trafficking Tools . . . . . 68
3.3.2 Positive Consequences of Anti-Trafficking Tools . . . . . 71
3.3.3 Negative Consequences of Anti-trafficking Systems . . . 74
3.3.4 Recommendations for Anti-trafficking Tool Users . . . . 83
**3.4 Proposed Framework for Developing Anti-Trafficking Tools: P3D Framework . . . . . 84**
3.4.1 Project Planning . . . . . 85
3.4.2 Data Preparation . . . . . 89
3.4.3 Design Phase . . . . . 93
3.4.4 Deployment Phase . . . . . 98
3.4.5 Summary . . . . . 100
**3.5 Discussion and Recommendations . . . . . 101**
3.5.1 Summary of Findings . . . . . 101
3.5.2 Responsible Research and Innovation Implications . . . 102
3.5.3 Recommendations for Policymakers . . . . . 103
**3.6 Conclusion . . . . . 104**
**3.7 Direction of the thesis . . . . . 104**

## 3.1 Introduction

This chapter addresses the first research question: *What are the challenges and potential unintended consequences of developing and using anti-trafficking software, and how can these challenges and negative consequences be mitigated?* By addressing this question, it highlights the challenges and risks associated with anti-trafficking tools. The aim is to use this information to develop appropriate measures to maximise the positive impact of these tools by proposing recommendations to mitigate negative unintended consequences.

This study uses the AREA framework to systematically address the research question by guiding its design and implementation. The AREA framework in RRI (UK Research and Innovation, 2023) provides a set of practices that encourage researchers to identify the intended and unintended consequences of research and innovation before the design stage and to devise strategies to mitigate any negative consequences. This aligns with the objectives of this chapter. Although predicting future outcomes of software tools is challenging, RRI encourages leveraging past experiences with such tools to help anticipate future outcomes (ORBIT, 2021). Consequently, this chapter evaluates existing anti-trafficking tools to understand the challenges involved in developing these tools and predicting their potential consequences.

This chapter focusses on Victim/Trafficker Identification tools, as they are the most commonly used, as discussed in Chapter 2. For the remainder of this chapter, the terms "anti-trafficking tools" and "anti-trafficking software" will be used interchangeably to refer to Victim/Trafficker Identification tools.

The AREA framework emphasises the importance of involving a diverse set of stakeholders in examining the implications of innovation (UK Research and Innovation, 2023; Pili, 2017). In accordance with this methodology, the stakeholder perspectives were incorporated through semi-structured interviews with the 13

participants. Through interviews and a review of existing literature, this chapter provides valuable contributions to the research community in four main ways:

1. It offers empirical insights into an underexplored area, highlighting both the positive and negative consequences and challenges of anti-trafficking tools.

2. It provides a set of design recommendations for stakeholders to consider when developing anti-trafficking software.

3. It demonstrates how the principles of RRI can guide the process of data collection and analysis, particularly when examining unintended consequences.

4. It presents a framework to help stakeholders develop anti-trafficking software that minimises unintended negative consequences.

This chapter also aims to use the findings, especially the framework, to guide the remainder of the thesis to help ensure that it minimises negative outcomes. In addition, it also elicits feedback from stakeholders engaged in this chapter on the thesis's plans to investigate and detect human trafficking recruitment in the US.

The remainder of this chapter is organised as follows. Section 3.2 describes the methodology used in this chapter, while Section 3.3 discusses the findings derived from stakeholder interviews. Section 3.4 introduces a framework based on these findings to assist stakeholders in the development of anti-trafficking software. This is followed by Section 3.5, which summarises the findings, examines the application of RRI within the Information and Communication Technology (ICT) domain, and offers recommendations for policymakers. Section 7.4.1 addresses the limitations of the research, and Section 3.6 concludes the study. Section 3.7 outlines the impact of the findings on the direction of the thesis.

## 3.2 Methodology

The Research Ethics Committee approved the use of interviews to carry out the work in this chapter under the reference CS_C1A_021_017_001 [Approval: 17-09-2021]. The ethics form used for the approval process is provided in Appendix

A.4. The following section discusses the concept of RRI and details how it was integrated as a key component of this study.

### 3.2.1 Responsible Research and Innovation

RRI plays a crucial role in addressing the potential risks associated with innovations before they are released to the public. However, there are currently no specific guidelines for implementing RRI in research or innovation projects (Jirotka et al., 2017). To address this issue, Stilgoe et al. (Stilgoe et al., 2013) developed a framework to support and implement RRI in the research process. This framework was later adapted and reworded by the UK Engineering and Physical Sciences Research Council, in collaboration with Stilgoe et al., and is now known as the AREA framework (UK Research and Innovation, 2023). The AREA framework aligns with RRI by informing study processes through the anticipation of potential risks and the mitigation of adverse outcomes. It does so by offering actionable steps to integrate responsible practices. The AREA framework is widely used in the UK (Yaghmaei and van de Poel, 2020) and comprises four key components (Pili, 2017; Owen, 2014). Each component of the AREA framework contributes to a comprehensive understanding of the ethical and social implications of research and innovation projects. The four components are as follows:

1. Anticipation - This component considers the possible consequences and impact of a research project or innovation (ORBIT, 2021; Stilgoe et al., 2013). This step involves examining how innovation can affect society, allowing the identification of challenges and risks (Pili, 2017). In addition, anticipation encourages the formulation of potential solutions to address these challenges and risks (ORBIT, 2021). It is crucial that the anticipation phase involves all relevant stakeholders to ensure that their perspectives and concerns are taken into account. Researchers typically engage with stakeholders through various methods, such as focus groups and interviews, to identify potential risks and implications associated with their projects (Pili, 2017).

2. Reflection - This component urges researchers to extend their focus beyond traditional considerations and critically evaluate the motivations and benefits of their projects (ORBIT, 2021; Stilgoe et al., 2013). This involves assessing whether research or innovation is necessary and whether its advantages outweigh potential risks. Researchers use reflection as an inward-focused approach to improve informed decision making, identify key stakeholders involved in the project, and encourage contemplation of broader questions related to the project at hand (Pili, 2017).

3. Engagement - This component involves engaging relevant stakeholders, such as experts and the public, in the research or project (Pili, 2017; Stilgoe et al., 2013), facilitating a wider range of perspectives by considering various points of view (UK Research and Innovation, 2023). Collaborating with a diverse group of individuals is an effective method of obtaining valuable insights, feedback, and guidance that can influence the direction of a project and the decision-making process (Pili, 2017).

4. Act - This component focusses on shaping the trajectory of research and innovation in a positive way by building on the insights gained from anticipation, reflection, and engagement (Pili, 2017). At this stage, stakeholder participation plays a crucial role, as their insights and recommendations help guide the ongoing design and implementation of the project (Pili, 2017).

This study incorporated these AREA components to achieve its research goals. Figure 3.1 presents an overview of how the AREA framework was integrated into the study, and further details are discussed in the following sections.

### 3.2.2 Anticipate: Analysing the Consequences of Anti-trafficking Tools

Anticipation in this study was achieved through interviews and a review of the existing literature. The interview questions were designed to gather participants' feedback on the possible consequences and challenges associated with the use and

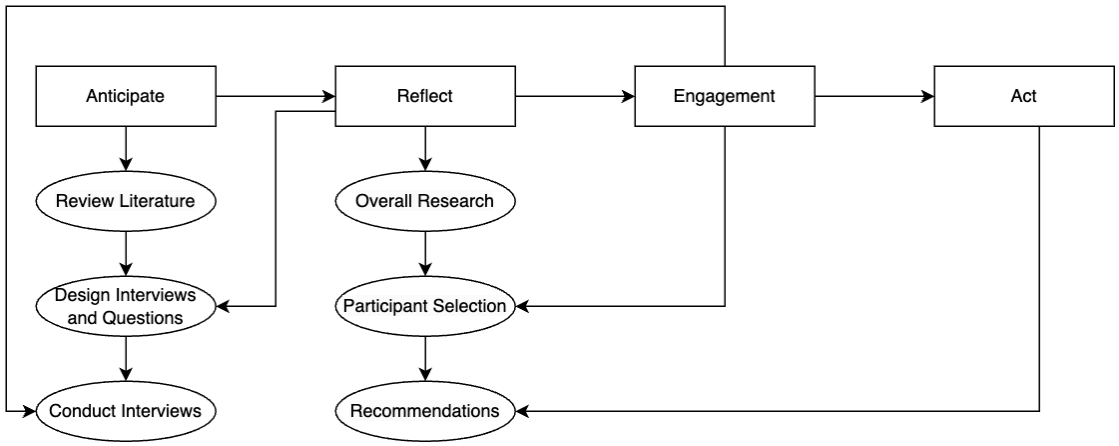


**Figure 3.1:** Chapter 3 Methodology Process

development of these tools, as well as their recommendations for mitigating negative implications. The interview questions can be found in Appendix A.1. Reviewing the existing literature provided essential background information on the topic and supplemented the insights gained from the interviews. It offered a deeper understanding of anti-trafficking tools, and their impact. The process of conducting the literature review was similar to that described in Chapter 2. Details about the interviews are discussed in Section 3.2.4.

### 3.2.3 Reflection: Examination of Motivations and Benefits

Reflection was integrated into the various stages of the study. Initially, it was used to assess the relevance and potential benefits of the study. In addition, reflection was used in this phase to guide the selection of key stakeholders for the interview, considering the diverse perspectives they could offer and the unique insights they could contribute. Due to concerns about the potential for (re) traumatisation, survivors of human trafficking were not directly included in the study as participants. Furthermore, there was a lack of support for survivors during the research process, in the event that welfare support was required. Instead, an alternative approach was identified to represent the voices of victims; this was achieved indirectly through participants with practical experience of working with them.

Reflection also played a crucial role in developing the interview questions, allowing me to carefully consider which questions would be most appropriate

for different groups of participants and which would yield the most valuable information for the research.

### 3.2.4 Engagement: Involving Diverse Stakeholders

In this study, engagement was integrated into the participant selection process to address the research question. The participants included law enforcement officers, anti-trafficking experts, and system developers from both the private sector and academia. These groups were chosen because law enforcement officers and experts working in non-governmental organisations are the primary stakeholders who use Victim/Trafficker Identification tools, while system developers are key contributors to the creation of these software tools (OSCE, 2020). The subsequent section provides additional information about how experts were engaged through interviews.

#### Recruitment and Interviews

**Recruitment** A total of 28 candidates were approached to participate in the study, of whom 13 were successfully recruited and interviewed. Data saturation was reached in the 12th interview, as no significant new codes emerged in the 13th interview. Saturation was determined through an iterative coding process, and since no additional themes were identified, recruitment was stopped. The study prioritised the acquisition of rich and detailed information from the diverse participants previously mentioned.

The recruitment process included various methods. Initially, three participants from existing professional networks were recruited by convenience sampling (Etikan Ilker et al., 2016). The snowball sampling (Naderifar et al., 2017) was then used to identify and recruit two more participants. In addition, a participant was obtained from an email list. The remaining individuals were selected by reviewing published work and LinkedIn profiles. Eight participants were approached by email; the remaining five were contacted through LinkedIn.

In particular, system developers were identified through academic papers and contacted using available contact information. LinkedIn was used as an alternative

if institutional emails were unavailable. Moreover, developers were identified by reaching out to the founders of companies known to develop anti-trafficking tools using LinkedIn. Law enforcement officers were recruited through existing networks and snowball sampling, with one officer identified through a LinkedIn search using keywords such as "law enforcement human trafficking" and "police human trafficking". Anti-trafficking experts were identified through a combination of existing networks, snowball sampling, and a review of published reports. The selection criteria for the participants included:

- System developers with experience in developing anti-trafficking tools.
- Law enforcement officers with experience in human trafficking investigations and anti-trafficking systems.
- Anti-trafficking experts with experience in the anti-trafficking field and knowledge on anti-trafficking tools.

The recruitment of participants in this study presented significant challenges, particularly in securing the cooperation of active law enforcement officials. As a result, the researcher extended the recruitment efforts to include retired law enforcement officers who had previous experience using anti-trafficking tools and maintained involvement in the anti-trafficking field through alternative channels. This recruitment and interview process lasted for six months.

Table 3.1 presents the characteristics of all participants, including their job status, technological background, and experience with anti-trafficking efforts. Participants are called "P" throughout the interview transcripts, followed by a numerical identifier (e.g., P01 designating Participant number 1).

**Semi-structured Interviews** Semi-structured interviews involve having a list of predetermined questions while allowing flexibility to explore other research areas based on participants' responses (Barclay, 2018). This type of interview was chosen for its flexibility, allowing me to probe further into responses and gain a more thorough understanding.

| Characteristics | Number | Percentage |
|---|---|---|
| Total Number of Participants | 13 | |
| Gender of Participants | | |
| Female | 6 | 46.2% |
| Male | 7 | 53.8% |
| Job Type | | |
| Anti-Trafficking Expert in org | 2 | 15.4% |
| Criminologist | 1 | 7.7% |
| Detective Constable | 1 | 7.7% |
| Retired Police Officers | 2 | 15.4% |
| Computer Scientists | 4 | 30.8% |
| Anti-Trafficking Expert/Founder | 3 | 23% |
| Work Experience | | |
| <5 Years | 3 | 23% |
| From 5 to 10 Years | 4 | 30.8% |
| >10 Years | 6 | 46.2% |

**Table 3.1:** Demographic Information of Participants

The process encompassed several key steps in the preparation for the interviews, including obtaining consent, scheduling interview times, and conducting pilot interviews. All eligible participants who met the criteria were contacted to arrange a suitable date and time for their online interview. Before the interviews, two pilot interviews were conducted to gauge the clarity of the research questions (Sampson Helen, 2004). Based on the feedback and results of these pilot interviews, necessary adjustments were made to refine the questions. The findings of the pilot interviews were not integrated into the analysis stage of the study.

The interview sessions took place from February 25, 2022, to July 21, 2022. These interviews were facilitated using Teams and a standalone recording device was used to capture audio recordings. The interviews ranged from 25 to 60 minutes, with an average duration of 45 minutes in all interviews. The interview format began with an introduction to the research.

Following this introduction, permission was obtained from the participants to initiate the interview and record it for documentation. The conversation started with general questions to gather information about the backgrounds of the participants. Subsequently, the interview discussed their experiences with anti-trafficking technologies and their perspectives on them. After these initial

questions, the interview proceeded into more detailed questions about the impact of the tools and their recommendations. Since all participants were native English speakers, the interviews were conducted in English. After the interview, the researcher transcribed the audio recordings and the subsequent analysis involved coding and analysing the data.

**Analysis**

Thematic analysis (Braun and Clarke, 2006), was used to extract meaningful information from the data. This method was chosen due to its flexibility and suitability for analysing data in underexplored areas (Braun and Clarke, 2006; Clarke and Braun, 2017). After transcribing the interviews, NVivo 12 Pro was used for the initial coding of all transcripts.

A total of 112 codes were identified and systematically organised into themes using axial and selective coding. As mentioned above, recruitment stopped after achieving data saturation. Although Braun et al. caution against relying solely on data saturation in thematic analysis, they advocate the consideration of contextual and practical factors, such as data richness (Braun et al., 2019). In this study, the research questions were carefully designed to elicit detailed data. In addition, the diversity of the participants contributed significantly to the richness of the data collected.

To ensure the reliability and credibility of the codes and themes, a triangulation method was used (Lauri, 2011; Jonsen and Jehn, 2009). Four participants were randomly selected and invited to provide feedback on the generated themes. This process aimed to assess whether their perspectives aligned with the analysis outcomes. The participants reached a consensus on the developed themes and their feedback also led to the identification of new information, which was subsequently incorporated into the existing themes.

### 3.2.5 Act: Integrating Stakeholder Perspectives

In this study, a framework was developed to assist anti-trafficking stakeholders in designing and implementing Victim/Trafficker Identification Tools that minimise potential harm as part of the Act process. The details of this framework are provided in Section 3.4. Additionally, participant feedback played a key role in shaping the direction of this thesis, a change discussed in Section 3.7.

### 3.2.6 Comparative Approach: US and UK Contexts

This study compared Victim/Trafficker Identification tools in the US and UK. As previously discussed, most of these tools have been developed in the US. The UK was selected for comparative analysis, particularly because certain tools developed in the US are also used in the UK. An examination of the tools in both countries revealed potential similarities and differences in the outcomes. This comparative approach facilitates a comprehensive understanding of the implications of these tools.

## 3.3 Research Findings

The analysis of the interviews identified four main themes: challenges associated with anti-trafficking tools, positive and negative consequences of anti-trafficking tools, and recommendations on the best ways to use anti-trafficking tools.

The themes are summarised in Table 3.2 below.

| Theme | Sub-theme |
|---|---|
| Challenges of Anti-Trafficking Tools | 1. Funding<br>2. Issues with technology companies |
| Positive Consequences of Anti-Trafficking Tools | 1. Distribution of resources<br>2. Aids in investigations<br>3. Prevents people from working with explicit content |
| Negative Consequences of Anti-Trafficking Tools | 1. Consequences of misclassifying traffickers<br>2. Victimisation of vulnerable people<br>3. Negative consequences for victims<br>4. Adversarial efforts of traffickers<br>5. False sense of security<br>6. Privacy implications of the tools |
| Recommendations for Anti-trafficking Tool Users | 1. Design and deployment recommendations |

**Table 3.2:** A Summary of Themes from the Interviews

### 3.3.1 Challenges Associated with Developing Anti-Trafficking Tools

This theme explores the challenges associated with the development and use of anti-trafficking tools.

#### 1. Funding

One of the significant challenges highlighted in this study was financial constraints. Anti-trafficking stakeholders often struggle with limited funds, which presents a major obstacle to developing and using anti-trafficking tools. In most cases, stakeholders must seek external assistance to develop these tools, which requires funding. Participant P07 highlights this issue by discussing the financial limitations faced by anti-trafficking stakeholders:

> "Most people working in the trafficking sector operate with limited resources. Many organisations and NGOs operate on a shoestring budget, making it difficult to allocate funds for large-scale technological advancement projects. As a result, investing in such initiatives, which could divert focus from front-line work, may not always be seen as a funding priority. (P07)

These financial limitations extend beyond the development stage. In some cases, additional resources are needed to acquire the infrastructure necessary to

support the operation of these tools and train personnel for their effective use. Moreover, the evolving landscape of human trafficking necessitates regular updates to these tools, which require further financial investment, presenting an ongoing challenge for anti-trafficking organisations.

Participants in the study, along with previous literature (Musto et al., 2020), noted that profit often drives companies to participate in the development of anti-trafficking tools. This profit-driven approach can have implications for the accessibility and affordability of these tools. Participant P06 highlighted the potential impact of profit-oriented companies on access to these critical resources:

> "There are organisations and private organisations that are trying to sort of probably make money out of this. And I don't think from my perspective, it's not about making money; it's about doing the right thing, sharing your technology, sharing your information, and making sure that you, in the first instance, and I'll stress this now, in the first instance, is to safeguard the victim." (P06)

Some participants further emphasised that funding represents a significant barrier to the prevalence of anti-trafficking tools in the Global South, particularly in regions like Africa. There are concerns about the scarcity of such tools in Africa and this was a recurring topic in interviews. Several countries in these regions face challenges in funding, infrastructure, and computer access, all of which are essential prerequisites to use these tools. This puts them at a disadvantage in combating trafficking compared to other developed countries.

### 2. Issues with Technology Companies

Participants expressed several concerns about the role of big technology companies in addressing human trafficking. These concerns directly affect how such companies support or fail to support stakeholders working to combat trafficking on their platforms.

A key issue raised was the lack of adequate regulation that governs the actions of these companies. The participants specifically pointed to the absence of laws that require content moderators to report suspected instances of human trafficking to the appropriate authorities. Although cases involving children often trigger mandatory reporting, there was concern about the inconsistent handling of cases involving

adults. This regulatory gap weakens accountability and, as a result, companies are often not incentivised to support stakeholders in developing detection tools for adult victims or to adopt existing anti-trafficking technologies on their platforms.

The participants also criticised the operational practices of big technology companies in relation to human trafficking. Despite growing evidence that traffickers use online platforms for recruitment and exploitation, these companies often fail to assist law enforcement during investigations, even when convicted cases of human trafficking originated on their platforms. With few legal obligations compelling cooperation, they are able to remain passive in the face of ongoing abuse. This lack of engagement contributes to the scarcity of ground-truth data, which in turn hampers the development and effectiveness of anti-trafficking tools, as fewer cases involving their platforms are successfully investigated and documented.

Participant P01 elaborated on the lack of accountability concerning these social media platforms, emphasising the challenges posed by this regulatory gap:

> "But you know, there is really no accountability, and these tools fundamentally hurt people. There is no question, it's a small percentage, but are they like responsible for that harm when people get hurt? Are they responsible for the user uploaded content that causes harm? So, one of the things that these companies just delete the content, are we safer as a society? Maybe a little bit but what about justice? Do the people get called to justice who were behind the accounts?" (P01)

The participants also highlighted that technology companies have a significant share of the data that could be instrumental in the fight against human trafficking. Despite possessing this valuable data, technology companies often withhold it, creating a significant disadvantage for efforts in this domain. Participant P12 addressed this issue, highlighting the crucial role these companies play in managing and providing access to the essential data needed to combat sex trafficking.

> "We have a human trafficking kill chain. And so, we would like to pay attention to each of the phases of trafficking and sort of ask the question, what does the trafficker need in order to achieve that portion of trafficking, right? And when you think through each of those nodes or each of those phases, you realise it's the private sector, far and away, who owns all the data for that, and who are the ones who have the strongest capability for doing something about it." (P12)

However, according to one of the participants, there has been a noticeable shift in recent years, with companies increasingly collaborating with anti-trafficking organisations. These partnerships aim to incorporate anti-trafficking detection software into their platforms.

### 3.3.2 Positive Consequences of Anti-Trafficking Tools

This theme explores the benefits of anti-trafficking tools for users, beneficiaries, and society. The findings in this section indicate that most of the positive outcomes were anticipated, aligning with the intended goals of these tools. However, some unintended positive consequences have also been observed and shown to be impactful. The impact of the tools is influenced by the country of deployment, as risks and benefits can vary depending on existing laws, practices, and culture of the country. These tools demonstrated similar positive effects in both the US and the UK.

#### 1. Significant Societal Changes

Participants highlighted that these tools offer a more comprehensive understanding of trafficking, facilitating the allocation of financial and non-financial resources to combat sex trafficking. These resources have notably benefited stakeholders, such as law enforcement and policymakers. For example, anti-trafficking tools have proven invaluable in pinpointing trafficking hotspots, allowing stakeholders to strategically channel their prevention and rescue efforts and make sound choices regarding resource distribution.

As instances of technology-facilitated trafficking continue to rise, identifying these hidden locations has become increasingly challenging. Anti-trafficking tools have proven helpful in addressing this challenge. Participant P02 shared their unit's experience in using these tools to locate victims:

> "We need to know where to look for them because they are just in normal houses and normal streets now in popup brothels, on Airbnb or short-term lets, so it is a lot more hidden. So, the adult service websites on the web we can do with that for the web scraping, means we could find where people are to try and safeguard them." (P02)

Furthermore, anti-trafficking tools have also served to draw attention to global challenges, notably by exposing the deficiencies inherent in existing legislation. Participant P01 discusses this issue and states:

> "But I'll just say the more we work these cases, the more we realise that the laws are imperfect, and so we have to figure out how we get the laws makers to make changes to the laws." (P01)

The awareness of these limitations has had positive repercussions. These tools have uncovered previously unknown information and provided valuable insights to stakeholders. This new knowledge has allowed stakeholders to identify areas in need of improvement, ultimately contributing to significant societal changes in the fight against sex trafficking.

For example, the participants highlighted that the intelligence provided by the tools played a pivotal role in the establishment of modern slavery units within the UK. These units have catalysed a shift in investigative approaches, from conventional policing methods to victim-centred policing. This transformation has proven to be particularly advantageous for victims. P01 discusses the impact of these tools on moving from traditional policing to victim-centred policing:

> "I think that having data-driven tools is an accelerator for transformation, so, like law enforcement and the public policy people, they have been working at their pace to try to move from traditional policing, victim centre policing, but once you have powerful tools that show the data, to point people like towards the resources and the direction that it needs to go. I think that's one way that our tool has been very helpful, to guide, just having a data-driven guidance to be able to accelerate the movements of victim centre policing." (P01)

### 2. Aids in Investigations

Most participants acknowledge that anti-trafficking tools have been instrumental in analysing the vast amount of data available on online platforms. These tools help gather information with just a click of a button, a process that would otherwise require significant time and resources.

These tools have benefited various stakeholders, particularly law enforcement agencies. The tools help prevent law enforcement from conducting sting operations

solely on the basis of intuition and without sufficient information. Intelligent operations ensure a more informed and evidence-based approach to their operations. Law enforcement agencies have relied on these tools to pinpoint trafficking locations and facilitate the rescue of victims.

In addition, the element of time is crucial for law enforcement agencies. These tools can save substantial hours of investigative time and valuable resources, which is essential given the time constraints and numerous commitments that law enforcement agencies often face. P03 elaborated on the benefits of these tools within the investigative process.

> "The information that they can gather with, with a touch of a button, compared to what that would take in terms of time, resource, etc, for probably numerous police officers over a period of months to get there is no comparison. And it's the way forward, and anyone who says different just doesn't know what they're talking about. I think that's the reality." (P03)

Furthermore, these tools have assisted law enforcement in gathering intelligence on broader criminal activities beyond sex trafficking. Human traffickers often engage in other illegal practices, such as drug trafficking. The tools inadvertently shed light on information related to these other criminal activities. P12 highlighted this unintended consequence as a positive outcome resulting from the use of the tools:

> "They are a valuable resource because they help build information and intelligence on wider criminal activities. If an organised network is involved in sex trafficking, for example, it is often polycriminal—engaging not only in sex trafficking but also in other crimes such as drug trafficking and money laundering. These groups are typically involved in multiple forms of criminality. This is where technology has played a role. In some cases, this may be considered an unintended consequence, as it has allowed for the development of broader intelligence on organised crime." (P12)

### 3. Prevents People from Working with Explicit Content

Anti-trafficking tools also help prevent stakeholders such as law enforcement officers and content moderators from directly engaging in explicit content. Working with sex trafficking content can be mentally and emotionally challenging, as individuals

expose themselves to explicit and potentially triggering material. The tools play a crucial role in addressing this challenge by automating processes and reducing the volume of explicit content that stakeholders would otherwise need to analyse manually. This reduction significantly eases the burden and improves the well-being of those working in this space. P12 acknowledged the difficulties associated with working in this environment.

> "Because you are, you know, for however many hours a day you have, you're looking at really sexually graphic content. And so, while it is not violent in nature, necessarily in terms of like, you know, we are looking at standalone pictures, we are not looking at video, but it is obviously very sexually graphic, there is a lot of a lot of the things that we are tagging are ethnic slurs, in terms of body descriptions, these different assessments or like language is really, you know, very sexually explicit certain acts that are being done. And so, when you're looking at these things through the lens of trafficking, it's really heavy content in terms of looking at it all day." (P12)

P04 also shares their experience working in this space:

> "So, I think those are kind of the, it's just so hard like I have empathy for anyone working in this field because it was so hard for me."(P04)

As stated above, these tools have been beneficial to a wide range of stakeholders. However, the following section discusses the negative consequences associated with them.

### 3.3.3 Negative Consequences of Anti-trafficking Systems

This section focusses on the negative consequences (intended and unintended) of the tools, as anticipated and discussed by the participants. They explore the challenges that may arise from using the tools. The findings in this section indicate that developers may sometimes not consider or foresee these negative consequences, but other stakeholders, with their experience in the field, can anticipate such issues.

### 1. Consequences of Misclassifying Traffickers

Participants expressed significant concerns about the high rate of misclassification associated with these tools, as they can have severe consequences for those wrongly identified as traffickers or victims. The interviews revealed two main causes for this misclassification: limited data availability and the inherent limitations of red flag tools.

**Limited Data -** Participants expressed concern about the scarcity of data in the anti-trafficking domain. A significant issue that has emerged is the need for ground-truth data to support technological solutions. Unfortunately, the lack of ground-truth data can be attributed to the low rate of prosecution in trafficking cases, which in turn impedes the acquisition of the required data for ML applications. There was a global decline of 42% in trafficking prosecutions between 2015 and 2018 (OSCE, 2020) and a decline of 27% in 2020 (United Nations: Office on Drugs and Crime, 2023). Previous research indicates that the decline in human trafficking prosecution can be attributed to the remarkably low risk faced by these criminals with respect to being identified, prosecuted, and sentenced (OSCE, 2020). Furthermore, traffickers' ability to operate without consequences further exacerbates the lack of reliable data. P8 discusses the issue of ground truth data and its critical role in developing tools using machine learning techniques:

> "In the trafficking space, applying machine learning is very challenging because ground-truth data is not always available. It is difficult to determine whether an advertisement corresponds to a trafficker, as prosecution is extremely rare, making it hard to obtain definitive labels." (P08)

The participants also discussed the underlying reasons for this low prosecution rate. In particular, they highlighted a trend in which courts often convict human traffickers for offences other than human trafficking. P10 provided information on this issue, shedding light on prosecution of trafficking cases.

> "So, you can't go into a massive database and do keyword searches or filter for human trafficking. Because the criminal activity in terms of the documents for which somebody is prosecuted might be different, so, we found very few actual cases of human trafficking prosecutions in the public domain and the public record, right." (P10)

The participants also expressed the lack of data sharing among stakeholders, worsening the issue of limited data availability. The reluctance to share data obstructs other stakeholders from accessing the necessary information to formulate effective anti-trafficking solutions.

Furthermore, participants linked the high rate of misclassification to the use of red flag tools or rule-based systems. These tools rely on predetermined indicators or patterns to identify potential trafficking cases. The following section discusses this further.

**Limitations of Red Flag Tools -** Previous research (Campana, 2022) raise concerns over the limitations of using indicators to identify trafficking, which several participants confirm as they state that red flag tools are associated with many false positives. These false positives can maintain a conducive environment for human traffickers to thrive. A participant expressed that the tools should only help provide information rather than rely solely on them to identify potential victims or traffickers.

One participant expressed complete scepticism about the effectiveness of red flag tools. P13 expressed doubts about the use of keywords or phrases by traffickers as reliable indicators:

> "I think there is zero probability that forced prostitution has the key as a keyword tell or some flag. Nobody advertises their illegality, nobody. And if it appears they are, it's likely that they are misleading. If I put an ad out, that says barely legal, right, or fresh meat, or something that's meant to indicate usefulness in the product I'm advertising. There is very, very little chance. So, I think almost no chance to indicate whether or not that person is over 18 or younger than 18." (P13)

However, some participants also acknowledged that software is not always the sole cause of these issues. In some cases, misinterpretation of the results by tool users can contribute to misclassification. The participants highlighted various

negative social consequences related to the misclassification of traffickers. They pointed out that mislabelling individuals as traffickers can result in stigmatisation and even lead to family breakdowns. P06, a law enforcement officer, spoke about the stigmatisation associated with such misclassification:

> "Now, sometimes they can get that wrong. And as we talked about the social consequences of that when you get that wrong, that stigma can be pretty heavy for the for the poor individual that's identified as being an offender that isn't so yeah, I mean, those consequences are pretty severe if you do get it wrong." (P06)

The participants also discuss that most of the offenders are males; therefore, families with heterosexual partners can leave women and children at a disadvantage. The following section discusses how misclassification can lead to the victimisation of vulnerable people.

## 2. Victimisation of Vulnerable People

Concerns about the potential victimisation of vulnerable individuals, particularly sex workers and people of colour, were often associated with the US, where sex work is illegal and there is a history of a turbulent relationship between police and marginalised communities. In such contexts, there is a high risk that tools contribute to discriminatory profiling and worsen existing inequalities and injustices.

Participants expressed concern that the tools, which often rely on personal information shared by sex workers on escort or adult websites, could lead police officers to approach sex workers based on this intelligence. They feared that this could result in their arrest or other negative consequences. One of the key challenges highlighted by the participants was the difficulty faced by tools or developers in differentiating between voluntary sex work and sex trafficking. This lack of distinction poses problems for sex workers, as it increases the risk of incorrect classification, potentially labelling them as victims when engaging in consensual sex work.

In countries where sex work is illegal, such as the US, this misidentification can also contribute to a criminal profile for voluntary sex workers. P04 talks about the possibility of arrest of the sex workers by stating:

> "Yeah, sex work is illegal. And buying sex is illegal in the US. So sometimes, if law enforcement has information, they might just arrest the sex workers who are actually being trafficked, rather than providing support and helping them have like a legal case against the trafficker if they want that, or just helping them get to someplace safe. Which is, you know, obviously terrible, because it's like, people don't need to be arrested and go to jail. Like, if they're experiencing trafficking, they need to, you know, like, have support." (P04)

Several participants expressed concerns about the difficulty in differentiating between sex work and sex trafficking. P10 raises moral questions that stem from the confusion between voluntary sex work and sex trafficking:

> "The other thing is if you look at if you look at the red flags, or the rules from the other companies that you have researched, as I look at those, I see that they confuse sex work with human trafficking. And I think as a society, we have to decide is that okay, or not? Right? Is it okay to treat all sex workers as if they're human trafficking victims? Because I don't think all sex workers want to be treated that way. And, I think that a lot of the interventions that I see the technologies confused the idea." (P10)

Additionally, P11 talks about how technology and individuals misidentify sexualised images as trafficking.

> "But over time, I'm very concerned about the way that technology is classifying things as trafficking that are not and it's been basically used like, the longer that the algorithms run, the more that that kind of gets impacted over time. And the way that, you know, images, or the sexualisation of certain images are going to be classified as trafficking and things are brought into kind of this loop. And I'm seeing that kind of in different parts of this work where people that are doing sex work, are being approached by law enforcement, and then counted as victims of trafficking." (P11)

Potential challenges can still arise for sex workers, even in countries where sex work is not illegal, such as the UK. Despite the illegality of sex work, some participants expressed concerns that law enforcement can still approach sex workers based on the information provided, which could have negative consequences.

Furthermore, some participants expressed concerns about the potential victimisation of people of colour, particularly in countries like the US, with a history

of strained relations between marginalised communities and law enforcement. They highlighted that individuals (mis)identified by the tools could face serious repercussions, particularly in contexts where racial profiling and discriminatory practices exist. P13 talks about not wanting to give police officers or other individuals tools that could be dangerous to people of colour:

> "And in America, if you stop a black man with a broken taillight, for having a broken taillight, you can probably determine that he might actually have some outstanding charge. And that might give you cause to push them to the ground, that might lead him to complain, which you can call resisting arrest, and then you can kill him. That's America. I don't want to give the police anything. So, if it's not the police in America, it's vigilantes, armed white evangelical vigilantes. I don't want to give them anything." (P12)

However, a law enforcement officer from the UK stated that law enforcement officers do not actively target people who are not victims. If this happens, the police must refrain from it and have protective measures against sex workers. There are also concerns about the negative consequences associated with actual victims of sex trafficking; this is described in the following section.

### 3. Negative Consequences for Victims

Participants reported that traffickers often force people to engage in illicit activities such as visa fraud by duress, coercion, or deception. It is possible for a nation's legislative framework and operational procedures to fall short of protecting victims, leaving them vulnerable to prosecution. In the UK, the 2015 Modern Slavery Act provides protection for victims forced to commit such crimes (Home Office, 2018. However, according to a review of the literature and some participants, the US does not have a similar law in place to protect victims. P01 emphasised the necessity for legal measures that prioritise the protection of victims:

> "There is some protections that need to happen for the victim. Victims are committing crimes as part of their exploitation, and we need indemnification for those crimes." (P01)

The criminalisation of victims can severely limit their opportunities for a better life after being rescued. Difficulties in reintegrating into everyday life, such as finding employment due to their criminal history, can discourage victims from viewing rescue operations as beneficial. Some may even resist rescue efforts if they perceive their current situation as preferable to their previous life. Unfortunately, these challenges can result in victims not wanting to be rescued, potentially leading them back into exploitation after leaving the trafficking ecosystem, thus perpetuating a cycle of victimisation.

In addition, some tools are specifically targeted at victims to help rescue them. However, these can be potentially harmful to the people they intend to help, as explained by P05:

> "Then the other thing is, what if someone, like if someone developed a tool that displayed something on the phone, such as example, like an ad for like a hotline, like the National Human Trafficking Hotline, and a trafficker saw that on the phone of someone who's experiencing trafficking, that could lead to violence, that is punishment, or possibly death, so by showing something that is, like an ad that's intended to help." (P05)

### 4. Adversarial Efforts of Traffickers

As law enforcement agencies investigate cases more efficiently, traffickers implement various adversarial tactics to avoid detection (Rivas et al., 2024). The OSCE highlights the displacement effect as a negative consequence of using anti-trafficking tools (OSCE, 2020). This effect refers to the concern that traffickers will disperse across various online platforms and frequently switch platforms as a tactic to evade detection.

Participants echoed these concerns and discussed other scenarios in which traffickers change their tactics to avoid anti-trafficking tools and stakeholders. For example, traffickers develop new creative ways to trick anti-trafficking tools, such as using emojis in their advertisements instead of text to avoid keyword searches or placing critical information on images to prevent detection from tools that do not use image detection. P09 speaks on their experiences with adversarial efforts by traffickers:

> "Another problem was that sometimes like the phone number was pretty important information. And sometimes the poster would just, instead of writing down the phone number on the text, they would put it on the image, or do something sort of like a CAPTCHA thing. So, it'd be very hard to call. So, you'd need like computer vision algorithms to call it which makes it way harder." (P09)

Furthermore, participants highlighted that traffickers deliberately omit critical information, such as phone numbers, from their advertisements to prevent the tools from identifying the locations of victims. These adversarial tactics pose significant challenges for technology solutions in detecting human trafficking.

The ever-changing tactics employed by traffickers can create fear and hesitation among anti-trafficking stakeholders in openly discussing or publishing their work. There is a concern that sharing too much information about the tools may inadvertently help traffickers understand how the tools operate, leading them to further adjust their behaviours to evade detection. Participant P08 specifically addressed this fear of publishing work in this domain:

> "I think first of all, when doing research in this space. There's not a whole lot of existing work going on, even though they're having a couple organisations or companies doing the anti-trafficking stuff, but the research is not often published. I think sometimes it's probably you know, those companies don't emphasise publication and the other one is that sometimes people deliberately don't share the research or data, in a fear that the traffickers might be adversarial, then they might also like read your methods and try to trick your methods down the road. So, there's not a whole lot of collaboration and research going on in that space." (P08)

Participants also expressed concerns about the relevance of the information used to develop anti-trafficking tools. As traffickers constantly change their tactics and switch to new platforms, specific keywords or phrases can become ineffective in identifying trafficking activities over time. Without regular maintenance and updates, these tools may lose their effectiveness in detecting and combating the evolving forms of trafficking. This degradation in performance and accuracy could give anti-trafficking stakeholders a false sense of security, which can result in other negative consequences.

## 5. Privacy Concerns

Several studies have raised concerns about the use of human trafficking data because the data is sensitive and the mismanagement of these data can lead to severe consequences (OSCE, 2020; Latonero and Movius, 2011; Latonero, 2012; Deeb-Swihart et al., 2019; Pendergrass, 2018; Rivas et al., 2024).

Participants echo these sentiments; they are concerned about the breach of privacy associated with using the tools. They are concerned about the tools collecting personal identifiable information that, if connected, could provide insights beyond the intended scope of data collection. Participants note that some people working in this field tend to gather personal information to identify victims and traffickers unnecessarily. P10 discusses the unnecessary need to collect personal data in this domain:

> "The second issue is that I see technologies violating people's privacy when there is no reasonable reason to justify collecting their personal data." (P10)

The participants also voiced concerns about the potential misuse of data collected by these tools, highlighting the alarming intelligence that the tools can derive from connecting the various data points. They expressed concerns about the invasive nature of these tools, particularly concerning sex workers.

As sex workers increasingly use online platforms such as adult/escort websites, these tools collect a large amount of information to identify potential instances of sex trafficking. These tools inadvertently collect data that are not related to human trafficking, such as personal data about sex workers (Deeb-Swihart, 2022). This information raises ethical questions about the extent to which they respect the privacy and consent of sex workers in the overall process. P11 discusses the ethical challenges of using escort ads as part of anti-trafficking work:

> "And with human trafficking, for so many years—and I include my own project in this—we have discussed what sex work looks like and the differences in interpreting a sex work ad. There is a distinction between someone willingly posting an ad in a specific way on a website and having that information ingested into a dataset for human trafficking analysis, which they did not consent to." (P11)

In addition to highlighting privacy concerns, some participants shared practices that can be incorporated to mitigate the potential negative impacts and challenges of these tools. The following section will explore these practices.

### 3.3.4 Recommendations for Anti-trafficking Tool Users

This section outlines recommendations for anti-trafficking stakeholders using the tools to help mitigate the negative consequences discussed in the previous section.

**Protocols -** Participants stressed the importance of setting up protocols for the tool's operation after deployment, such as establishing the schedule for software reviews and maintenance, which is a shared responsibility between users and developers. These protocols are essential to ensure the ongoing effectiveness and dependability of the tool, thus promoting its long-term sustainability and relevance in the fight against human trafficking.

**Human-in-the-Loop -** Participants advocated for the inclusion of human participation in the decision-making process when using these tools to ensure the accuracy of the information obtained. Experts should evaluate the outcomes generated by the tools and make final decisions, rather than relying on them as completely accurate. This is particularly important because these tools are prone to producing false positives, underscoring the need for explainability in the decision-making process.

**Address Privacy Concerns -** A law enforcement officer (P02) recommended that law enforcement agencies restrict access to the tool and its data to only relevant individuals. This approach should also be applied to other anti-trafficking entities. Implementing access controls is essential to prevent misuse of the tools and ensure security. Physical and logical access controls can be considered to limit access to servers or machines that store the tools, with additional restrictions on tool access to further safeguard sensitive data (Wright, n.d.).

The participants also recommended deleting unnecessary personal information. This approach would contribute to preventing the retention of data pertaining to vulnerable individuals, such as sex workers, and mitigating the probability of adverse consequences, including the misuse of data for purposes not related to trafficking.

The participants also provided several recommendations for developers when developing tools, which are outlined as part of a framework in the following section.

## 3.4 Proposed Framework for Developing Anti-Trafficking Tools: P3D Framework

This section presents a framework for developing anti-trafficking tools, termed the P3D Framework, which consists of four main phases: project planning, data planning, design, and deployment. The P3D Framework was developed to help anti-trafficking stakeholders address key issues and mitigate negative unintended consequences throughout the process of planning, designing, and deploying anti-trafficking tools.

The framework's components were developed through thematic analysis of participant interviews, as described earlier. Recurring themes from their recommendations on anti-trafficking tool development were mapped onto the four phases of the P3D Framework. These empirical findings, supplemented by a literature review, provided domain-specific recommendations to address the unique challenges and risks in designing anti-trafficking solutions.

The P3D Framework builds on and aligns with the core components of the AREA framework. Anticipation is embedded in the project planning phase to identify potential risks and ethical concerns early. Reflection spans all four phases, ensuring a continuous assessment of ethical and practical considerations. Engagement is prioritised by providing strategies for involving diverse stakeholders, while action is integrated into the design and deployment recommendations in the final two components.

By incorporating the principles of the AREA framework, the P3D Framework reflects RRI values while being tailored specifically to anti-trafficking tools. It translates broad RRI principles into actionable steps, serving as a domain-specific

guide that helps developers address real-world challenges in an effective and ethical manner.

The visual representation of the framework is shown in Figure 3.2. The following sections provide additional details on each of the main phases of the framework.

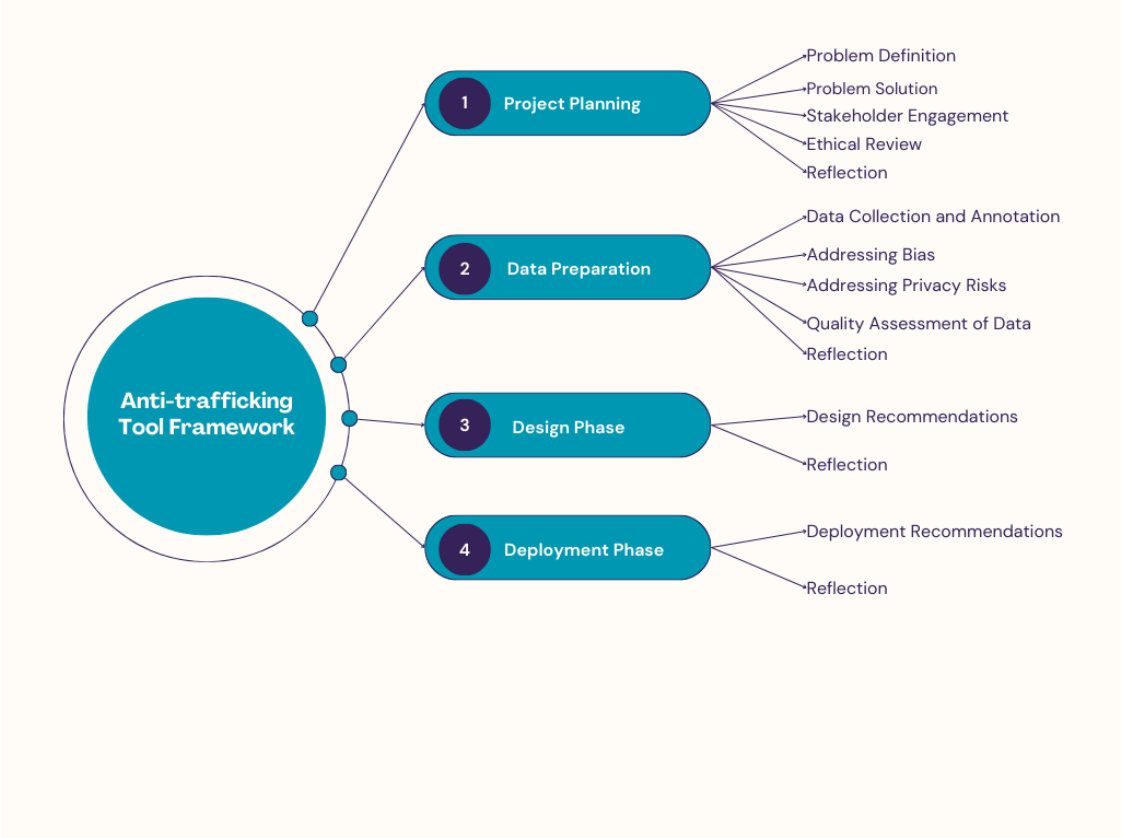


**Figure 3.2:** P3D Framework

### 3.4.1 Project Planning

This part of the framework establishes the groundwork for the tool by ensuring that the problem, the proposed solution, and the objectives are clearly defined. It also discusses strategies for identifying stakeholders for collaboration and key considerations for conducting an ethical review of the project before development.

#### Problem Definition

Developing an anti-trafficking tool requires a clear understanding of the specific problems being addressed (Deepak, 2023). Human trafficking is a complex issue

that takes various forms, each involving distinct practices, although its main stages are often similar. Developers should define the exact type of trafficking they will focus on. The participants emphasised the importance of thoroughly understanding the problem that the tool aims to solve. Therefore, system developers should conduct comprehensive research and analyse the current literature, private sector projects, and reports, such as the United Nations Office on Drugs and Crime's annual Global Report on Trafficking in Persons to identify information gaps and refine the problem area. System developers can also use qualitative methods to identify issues within the domain. Techniques, such as interviews and focus groups, are helpful in uncovering problems and unmet needs among stakeholders. Based on the availability of resources, stakeholders can employ multiple methods to address the problem at hand.

**Problem Solution and Objectives**

After identifying the problem to be addressed, stakeholders can develop a tool to fill this gap. The proposed tools should fall into one of the following categories: prevention, identification, or protection (OSCE, 2020). According to OSCE, classifying a tool into one of these categories helps users quickly understand its purpose (OSCE, 2020). Developers should ensure the following:

- Provide a clear description of how the tool will address the problem (Deepak, 2023).
- Ensure that the primary beneficiaries of the tool are victims or potential victims and design the tool with their needs in mind.
- Be transparent about the areas the tool addresses and its limitations.
- Determine the location(s) where the tool will be deployed and ensure that they have sufficient contextual understanding of these locations.
- Prepare detailed documentation of the problem and its solutions, as clear communication is crucial to managing the expectations of both users and beneficiaries.

### Stakeholder Engagement

As previously mentioned, discussing the plans for the proposed solutions with various experts is essential to ensure their practicality and usefulness. However, identifying these stakeholders is a significant challenge. To assist system developers in academia and the private sector, who may encounter difficulties accessing stakeholders, the following methods are recommended:

1. Use existing networks to identify relevant individuals or organisations.

2. Subscribe to mailing lists associated with initiatives such as Rephrain (https://www.rephrain.ac.uk/), which focusses on using technology to combat online harms such as human trafficking. The dissemination of requests through these mailing lists can help establish contact with potential stakeholders.

3. Conduct a systematic review of the scholarly literature to identify researchers who have previously performed work in areas of interest or possess relevant expertise.

4. Conduct internet-based searches for anti-trafficking organisations and initiate contact with them using the communication methods provided on their respective websites.

5. Use LinkedIn to find people with relevant experience or in relevant roles. These individuals can be identified by applying keyword searches.

6. Establish communication with founders and employees of organisations known to use technology to detect trafficking.

It is also crucial to involve stakeholders in discussions about the ethical implications of the project prior to the design stage. The following section discusses the ethical considerations that must be considered.

### RRI Review

Human trafficking is a sensitive area that requires careful consideration. Any errors or missteps in this area can have severe consequences (Deeb-Swihart et al., 2022). It is necessary that the project undergo an in-depth RRI review process, conducted by both domain and non-domain experts, to evaluate the project's decisions and strategies. This review process will help to make informed decisions about the project and evaluate the proposed strategies.

During this RRI review process, stakeholders are required to maintain transparency and openness about the project objectives, intentions, anticipated outcomes, and any potential positive or negative implications of the proposed solution. Furthermore, the review should evaluate other crucial factors, such as the suitability of stakeholders and the availability of resources to support survivors if they are included. This assessment facilitates the alignment of project goals with ethical standards and ensures that all parties, especially survivors, are adequately supported.

### Reflection: Overall Questions

After concluding project planning, stakeholders should consider and reflect on the following questions before proceeding to the data planning and tool design stages. It is essential that stakeholders have responses to these questions and that the answers are clear and well-clarified.

- What is the specific problem the tool aims to address?
- What type of trafficking does the tool focus on?
- Is the tool tackling in-person or online trafficking, or both?
- How does the tool address existing gaps in combating trafficking?
- What are the stakeholder perspectives on the proposed solution?

### 3.4.2 Data Preparation

This aspect of the framework guides stakeholders in considering various factors that can improve data quality during collection, annotation, and preprocessing. Since data is a crucial determinant of algorithmic performance, high-quality data will lead to better results (Kniazieva, 2022). A well-documented plan detailing how stakeholders will prepare their data, taking into account the factors discussed in this section, would be valuable for review by other parties.

#### Data Collection and Annotation

**Data collection -** This aspect is critical to the creation of the tool, as it directly impacts its performance (Kniazieva, 2022). System developers may sometimes use existing datasets, but when custom datasets are needed, it is essential to develop them effectively. The quality of the data is a key factor in the success or failure of the tool, as highlighted by the negative consequences discussed in Section 3.3.3.

When using AI as a component of the solution, it is imperative to establish an understanding of the AI problem statement before initiating data collection (Deepak, 2023). The AI problem statement can be derived from the problem definition in the previous section. This statement helps to ensure that the collected data are pertinent to the issue being addressed and that an appropriate solution will be developed (O'Regan, n.d.).

Following data collection, system developers should assess the volume and adequacy of the data. In cases where data are insufficient, it is advisable to explore secondary sources or collaborate with other stakeholders to obtain additional data. Developers may also consider using synthetic data, though this approach requires rigorous testing and validation. Furthermore, techniques such as data augmentation can be used to enhance datasets (Mumuni and Mumuni, 2022).

**Data Annotation -** In the field of online human trafficking, stakeholders should aim to perform data annotation with the help of experts. Stakeholders

must collaborate with experts who possess the necessary domain knowledge and experience to facilitate the annotation process.

Data annotation is a significant contributor to data bias, which arises from various factors, including race and gender (Paullada et al., 2021; Hube et al., 2019; Monfort et al., 2020. Previous research (Paullada et al., 2021) highlights the interpretive nature of data annotation and the absence of ground truth presents a major issue in verifying the data. Strategies to mitigate bias and identify ways to help verify data are discussed in the following section.

### Address Bias

Data bias occurs when training data contain superficial cues or features associated with labels that are not relevant to predict real-world scenarios (He et al., 2019; Teney et al., 2020). This bias can result in negative consequences and inequitable outcomes, particularly for marginalised communities and potential victims, as discussed in Section 3.3.3.

To mitigate data bias, stakeholders should establish a clear plan. Several approaches should be considered when developing such a plan.

- Participants and previous research (Paullada et al., 2021; Seneor and Mezzanotte, 2022) recommend developing improved data collection and annotation methods. It is suggested that the teams responsible for data collection are diverse, including representations based on demographics and skill sets. Furthermore, stakeholders must ensure that the data itself are diverse and adequate (Seneor and Mezzanotte, 2022. Machine learning models are known to perform poorly for people of colour and women (McKenna, n.d.). Ensuring that the data are diverse will help avoid these issues.

- To address the issue of annotation bias, system developers are recommended to involve at least one expert in the data labelling process to perform consistency and validation checks on the annotations. In addition, it is important to subject the dataset to iterative refinements to improve the quality of the labels.

- System developers are encouraged to use algorithms and tools to eliminate bias in datasets using debiasing strategies. These algorithms and tools (He et al., 2019; McKenna, n.d.; Bolukbasi et al., 2016; Bellamy et al., 2018), such as DRiFt (Debias by Residual Fitting), employ statistical techniques to detect and mitigate biases.

  Most data bias strategies involve the use of statistical techniques; however, Paullada et al. (Paullada et al., 2021) recommend that researchers also exercise their judgment to verify these inaccuracies. It is important to evaluate the consequences of the dataset when using algorithms, as the removal process can lead to additional problems, lower accuracy, and affect the decision making process of a model (Khani and Liang, 2020).

**Addressing Privacy**

Stakeholders must be transparent about their plans to protect user data during the collection, processing, and storage stages. To help ensure privacy, stakeholders should consider the following aspects when developing the tool:

- System developers should prioritise implementing data minimisation. This involves collecting only data that are essential to achieve the objectives (Goldsteen et al., 2022). Specific data requirements may differ according to the type of anti-trafficking tool that is being implemented. When only relevant data are collected, the risk to victims and other susceptible groups in society is reduced. In addition, it simplifies the process of ensuring data privacy (DataGrail, 2023).

- Anonymisation and pseudonymisation should also be considered because they are helpful when handling sensitive data related to human trafficking, especially when such data are not public. These techniques prevent the potential identification of individuals. These techniques are employed in various fields, one of which is the healthcare sector, which uses them to protect patient data (Al-Zubaidie et al., 2019; Langarizadeh et al., 2018).

- Developers should follow secure coding practices when developing tools. This involves performing routine code reviews, assessing vulnerabilities, and adhering to best security practices. One participant recommended that developers should use separate computers for data scraping and processing to provide additional protection for data.

- Various machine learning techniques and frameworks can strengthen the security of anti-trafficking systems. For example, privacy-preserving models can be implemented, differential privacy frameworks integrated, and noise added to data to prevent the extraction of sensitive information (Papernot et al., 2016). Furthermore, machine learning methodologies can facilitate the adjustment of privacy settings after implementation as data flows evolve within the system (Anthonysamy et al., 2017).

The list provided above is not exhaustive, and it is essential that stakeholders continue to explore additional ways to ensure data privacy. Other privacy measures have been discussed above in Section 3.3.4.

### Quality Assessment of Dataset

Several methods are available to improve the quality of datasets to help comply with industry and government regulations. These methods can be divided into automated and manual inspection.

**Automated Inspection** Automated data inspection involves a thorough examination of a dataset to identify and correct issues such as spurious correlations and data bias (Paullada et al., 2021). Various automated methods (He et al., 2019; Swayamdipta et al., 2020; A. Wang, 2020; Sakaguchi et al., 2019) are available to assess datasets for these issues, which can accelerate the evaluation of data quality in terms of accuracy, consistency, and usability, while also reducing the potential for human error (Egnyte, 2021). However, stakeholders should carefully consider the risks of employing current or custom-built automated methods to evaluate datasets.

**Manual Inspection** Manual inspection of data is useful for assessing quality and identifying unexpected or harmful content (Paullada et al., 2021). Although automated dataset assessments are beneficial, manual inspections can uncover additional issues that may be overlooked. They also provide valuable insights on improving data quality, such as detecting subtle errors, inconsistencies, and other problems that automated methods may miss (Kreutzer et al., 2022).

**Reflection: Overall Questions**

The following questions must be reflected on or addressed before moving on to the design phase of the framework.

- What is the rationale for collecting these data?
- What is the adequacy of the data in terms of suitability, diversity, and sufficiency?
- Who will have access to these data?
- How can these data be used for other unintended purposes?
- What are the strategies to maintain the relevance of the data?
- What will happen to the data over the next five years? How will you secure the data?

### 3.4.3 Design Phase

The development of anti-trafficking tools requires careful consideration during the design phase, as it can prevent many negative consequences that arise when using these tools. When developing anti-trafficking tools, it is essential to consider the following design considerations.

**Enhance Collaboration and Work with Experts -** The participants emphasised the importance of involving individuals with extensive knowledge of human trafficking during the early stages of tool development. They suggested forming a specialised working group of experts that also includes sex workers and survivors, with advocates representing them when necessary. Survivors, defined as individuals who have successfully escaped trafficking, provide invaluable information, especially those with expertise in technology-facilitated trafficking. Although opinions varied on the extent of survivors' participation in the development process, most participants agreed that their involvement is crucial during the design phase to identify potentially harmful elements.

Previous research (Deeb-Swihart et al., 2022) has highlighted the importance of involving survivors in vetting the ground-truth data used to train AI models, thus enhancing the accuracy of these tools. Methods such as focus groups can effectively engage survivors in the development process. Deeb-Swihart further recommended that survivors and disadvantaged communities should have a sense of ownership in the project, whether as project leaders or researchers, to ensure equitable participation (Deeb-Swihart et al., 2022).

Furthermore, given the ever-evolving nature of human trafficking, expert input is vital to confirm the feasibility and suitability of proposed tools. The participants recommended that diverse teams be involved to thoroughly examine the tools prior to deployment. Such collaboration would help ensure the practicality and effectiveness of the tool.

**Consider the Deployment Context -** Developers should thoroughly investigate the deployment environment during the design phase, as the impact of a tool is heavily influenced by the legal framework, cultural norms and operational practices of its location. Developers should customise the tool to the specific context, as certain features may produce unintended negative consequences in different settings. Taking into account these contextual factors, developers can tailor the tool for optimal performance in each deployment environment.

**Simulate the Academic Style of Working -** Several participants suggested that the private sector should adopt practices commonly used in academia. Academics are renowned for their extensive documentation of their methodology, intentions, and justifications for the problem and for building on previous research. Establishing clarity in the plans from the outset is crucial for developers. This clarity should encompass the methodology they intend to employ, their underlying intentions, and the proposed solutions they intend to implement. Previous studies (Musto et al., 2020) and participants have indicated that the motivations for developing anti-trafficking tools are not always clear, and adopting an academic writing style helps ensure that the research problem has been thoroughly explored and the contributions are clear.

**Increase Explainability -** Both previous research (Justice and Home Affairs Committee, 2022; Annah Whyton, 2014; Burkart and Huber, 2021; Freitas, 2014) and study participants have highlighted the importance of explainability regarding tools such as anti-trafficking tools. Transparency and explainability are essential to facilitate informed decision-making among experts and to assess the reliability of algorithmic results.

The United Kingdom's Justice and Home Affairs Committee has raised concerns about AI systems that predict crime, particularly about the potential for unfair criminalisation based on opaque algorithms (Justice and Home Affairs Committee, 2022). Understanding the reasoning behind algorithmic decisions is crucial for building trust, which is vital for both experts and society, especially when these insights are used in legal defences to justify decisions (Annah Whyton, 2014; Burkart and Huber, 2021; Freitas, 2014).

However, achieving explainability can be challenging, especially with complex models that are not inherently interpretable (Annah Whyton, 2014; Burkart and Huber, 2021; Freitas, 2014). Therefore, it is essential to enhance the interpretability of models used in anti-trafficking software.

Post hoc explanation techniques, such as LIME and SHAP, offer a way to interpret otherwise opaque models. These techniques involve developing additional models to explain the results of the initial models, thereby increasing trust, promoting fairness, and mitigating potential harms such as bias (Maveli, n.d.). Stakeholders can leverage multiple explanation models to improve confidence in the results (Salih et al., n.d.; Vimbi et al., 2024). However, while post hoc explainable models offer significant benefits, caution is advised when applying them in sensitive areas (Hassija et al., 2024).

**Strengthen Adversarial Efforts Handling** Participants strongly encourage developers to create solutions that counter the adversarial efforts of traffickers. To effectively counter the evolving strategies of human traffickers, it is crucial to adopt a proactive and adaptable approach that incorporates advanced methods such as adversarial training, transfer learning, and ensemble learning (Bai et al., 2021; Zhuang et al., 2021; Donges, 2019). In addition, developers can address the changing tactics of traffickers by integrating feedback loops. These feedback loops refine model elements after deployment using model output or user feedback (Soni, 2022). This approach allows the model to learn from new data provided by experts and adjust hyperparameters, allowing it to adapt to new challenges and evolving tactics (Soni, 2022; C3.AI, n.d.; Wu et al., 2022). It is essential to determine the most appropriate approaches for the anti-trafficking tool early in the development process.

**Conduct Algorithm Auditing -** Algorithm auditing involves evaluating performance and identifying potential societal harms, such as biases, as revealed in (Justice and Home Affairs Committee, 2022). Independent evaluations conducted by auditors provide valuable information on internal processes and documentation, thus improving accountability and transparency (Metaxa and Hancock, 2022; Digital Regulation Cooperation Forum, 2022). These auditors should be independent third parties. Algorithm audits address some of the negative consequences raised in the previous section, such as bias and victimisation of vulnerable victims. Metaxa and Hannock proposed an algorithm to audit tools with significant social impact

(Metaxa and Hancock, 2022). However, they also recommended anticipating the consequences of such audits and addressing potential issues that may arise (Metaxa and Hancock, 2022). The AREA framework (Stilgoe et al., 2013) and the AREA Plus framework (Jirotka et al., 2017) can be used to assess the impact of these audits.

**Implement Audit Trails for Supervision -** Incorporating an audit trail into anti-trafficking tools is beneficial for ensuring accountability and preventing misuse of the tools. Anti-trafficking tools have the potential to be misused in contexts such as stalking ex-partners. To mitigate these risks, participants recommended adopting a zero-trust mindset and actively monitoring the use of these tools. Stakeholders should conduct regular audits of tools and their data handling processes to ensure ongoing compliance with privacy standards and identify and address vulnerabilities. These audits also offer an opportunity to identify any weaknesses or potential risks associated with the tools.

**Incorporate Human-in-the-Loop Approaches -** The Human-in-the-Loop (HITL) approach emphasises the active role of humans in the development process, including data collection, model training, evaluation, and optimisation. This technique allows human intervention in the learning and decision-making processes of a machine learning system by providing input and guidance (Wu et al., 2022; Mosqueira-Rey et al., 2023; Xin et al., 2018).

The participants recommended adopting the HITL approach in anti-trafficking software to improve accuracy. Involving domain experts in the decision-making process can also help reduce false positives, preventing potentially serious consequences. Furthermore, the HITL approach improves the clarity and interpretability of the model results, particularly in natural language processing tasks (Wu et al., 2022).

**Ensure Ease of Use -** Participants and previous research have highlighted the need for user-friendly tools (Campana, 2022). Developers must design tools that cater to diverse user bases with varying levels of technical proficiency. Simplified tools can help reduce the risk of misinterpretation and cater to people who do not have

experience with technology. The law enforcement officers in the study emphasised the importance of ease of use, as most of the officers were not technically proficient. The development of user interfaces that facilitate a seamless information flow can help address this issue. Developers are recommended to incorporate user experience (UX) frameworks or principles, such as the UX honeycomb (UXPIN, n.d.), to ensure positive UX. Collaborating with anti-trafficking stakeholders during development ensures that users' needs, values, limitations, and abilities are understood.

**Reflection: Overall Questions**

- What strategies are employed to ensure the continued relevance of the tool?
- How is the tool optimised to search for pertinent information effectively?
- Who will use the tool?
- What are the key interests of the users regarding the tool's functionalities?
- In what ways could different stakeholders use the tool, intentionally or unintentionally?
- Who is affected by the algorithm's decisions?
- How transparent are the algorithmic decisions?
- How can the potential misuse or unintended use of the tool be mitigated?

### 3.4.4 Deployment Phase

A strategy for deploying anti-trafficking tools is necessary, as there are important elements to consider to ensure that AI tools perform as designed throughout their use and produce high-quality results. The deployment considerations are summarised below.

**Assessment of Infrastructure -** It is crucial for anti-trafficking stakeholders to assess their existing infrastructure to ensure that the tools will effectively support their operations. Some stakeholders operate with limited financial resources, and implementing a tool that requires ongoing maintenance can strain their budgets. This evaluation should determine whether the current infrastructure is adequate for both the use and maintenance of anti-trafficking tools. System developers should assist stakeholders in accurately estimating the financial resources needed for both the short- and long-term use of the tool.

**Use of Custom User Tests -** Campana highlights the importance of subjecting tools to comprehensive testing prior to deployment, a principle consistent with established practices in the field of ICT (Campana, 2022). However, participants suggested that system developers perform additional custom evaluations to assess the effectiveness and impact of the tools. This process must be carried out in collaboration with anti-traffickers stakeholders. Such assessments are critical steps to ensure that the tools meet their intended objectives and align with the design recommendations suggested above.

**Establish Clear User Agreements -** Participants proposed developing user agreements that define the acceptable use of tools and stipulate any limitations on their use. This is a crucial step in establishing clear guidelines for the proper use of the tools. These guidelines would also create a legal framework to discourage misuse and outline both legal and non-legal consequences that could arise from such misuse. For example, developers could reserve the right to terminate access to the tool if misuse is detected, as outlined in the user agreement.

**Provide Effective User Training -** Previous literature (Campana, 2022) and participants emphasised the importance of restricting the use of anti-trafficking tools to qualified and skilled individuals. Thorough training is essential to ensure that experts can use these tools accurately, thereby avoiding the misinterpretation issues mentioned in the previous section.

In addition, it is crucial to train a sufficient number of experts to ensure continuity in the use of these tools, even if trained individuals leave the organisation. This approach helps maintain the effectiveness of anti-trafficking tools despite personnel changes. The participants and previous literature (Freeman et al., 2022) have highlighted that there are high rates of transfer and retirement among police officers. Therefore, law enforcement agencies could benefit from training a substantial number of officers.

**Reflection: Overall Questions**

- What strategies will be implemented to effectively introduce the tool in the designated country, city, or state?
- How will false positives be identified and addressed within the framework of the tool's operation?

### 3.4.5 Summary

While many design guidelines exist, they often focus on general software development principles and fail to address the specific challenges involved in developing anti-trafficking tools, particularly the relevant ethical considerations. The P3D Framework was created by drawing on insights from experts with experience in human trafficking and anti-trafficking technologies, aiming to identify strategies for reducing risks during tool development. Through expert recommendations and reflections, developers can better account for elements unique to anti-trafficking efforts, helping to ensure that ethical, social, and technical challenges are effectively addressed.

Although the framework is presented in a top-down manner, it does not need to be followed in a strict sequential order. It is designed to be adaptable to various modern software development methodologies, including agile development, which has become a widely used approach in recent years. Developers can prioritise and iterate on certain stages or elements of the framework, such as stakeholder engagement, ethical reflection, and data planning, through multiple agile sprints based on their project goals.

It is recommended to incorporate all elements of the framework to maximise impact and mitigate risks; however, developers have the flexibility to integrate parts of the framework into their existing workflows. The framework serves as an initial guide, outlining key considerations in anti-trafficking software development, but it is meant to be adapted and refined according to specific project needs.

## 3.5 Discussion and Recommendations

### 3.5.1 Summary of Findings

This study assessed the challenges and potential consequences associated with the use and development of Victim / Trafficker Identification Tools in the US and the UK. It also offered recommendations to help mitigate these challenges and potential negative outcomes. It used the AREA framework to guide the activities of the study.

The study produced five main themes: challenges in developing and using anti-trafficking tools, positive and negative consequences of anti-trafficking tools, and recommendations for users and system developers when developing and using these tools.

Anti-trafficking tools have proven their value by assisting law enforcement agencies in trafficking investigations. They provide information on additional criminal activities associated with traffickers and significantly reduce the manual analysis of data online. In addition, these tools have the potential to bring about societal change by raising awareness and enabling proactive measures against human trafficking.

However, this study also sheds light on the negative consequences and challenges that can accompany the use of these tools. These drawbacks encompass concerns related to privacy breaches, potential harm to victims, countermeasures employed by traffickers, adverse effects on vulnerable populations, funding, lack of regulations for big tech, and the social repercussions of misclassifying individuals as traffickers. Furthermore, challenges exist in securing funding and obtaining support from technology companies during the development and implementation of anti-trafficking tools.

To address these negative consequences, study participants have suggested a series of considerations for developers and users of anti-trafficking tools, such as explainability, collaboration with experts, and human-in-the-loop approach. These considerations improve the effectiveness, ethical standing, and accountability of anti-trafficking tools. They highlight the importance of a responsible approach to designing and developing these tools considering possible societal impacts.

Finally, this study presents a framework to assist anti-trafficking stakeholders in the development and implementation of responsible anti-trafficking tools.

### 3.5.2 Responsible Research and Innovation Implications

The AREA framework in RRI is designed to be a useful tool for researchers and innovators in various fields. This framework aims to better align the results of research and innovation projects with the needs and values of society. By incorporating the principles of RRI into the design and execution of a project, researchers and innovators can engage with a diverse range of stakeholders in an open and transparent way, thus improving the acceptability of their work by society as a whole (UK Research and Innovation, 2023).

In the ICT space, components of the AREA framework can help anticipate innovation outcomes early in a project. Assessing existing relevant systems, as demonstrated in this chapter, can provide a deeper understanding of the potential consequences of proposed technologies. However, future work can consider incorporating quantitative methods into the research methodology to offer a more holistic view of the anticipated consequences (Shorten and Smith, 2017; Creswell J. W. and Plano Clark, 2006).

The AREA framework also encourages continuous reflection throughout the project. Although it is typically applied at the beginning of a research project to determine whether to proceed (ORBIT, 2021; Pili, 2017), it can also be used to examine various aspects of the research, such as interview participants, findings, and methods, as was done in this work. In doing so, researchers can gain critical insight that can change the direction of their projects.

In addition, the framework recommends that researchers consult diverse groups of experts when assessing the potential impact of their research projects or innovations. By engaging with various groups of experts, researchers can gain a deeper understanding of potential problems that may not always be immediately apparent (UK Research and Innovation, 2023). This is especially crucial in projects with social implications, as technical experts may not always fully comprehend the social consequences of their work; this is evident in the present study.

The implementation of RRI in practical applications can pose challenges to researchers and innovators due to the lack of explicit instructions provided by RRI and its accompanying framework, AREA. Furthermore, the unpredictable nature of the risks associated with ICT and the complexity of projects involving multiple stakeholders make the implementation of RRI particularly challenging in the ICT sector (Jirotka et al., 2017; Von Schomberg, 2013). To address these challenges, Jirotka et al. introduced the AREA Plus framework, which was specifically adapted for ICT projects (Jirotka et al., 2017). Researchers who encounter difficulties in implementing the AREA framework can find the AREA Plus framework beneficial.

### 3.5.3 Recommendations for Policymakers

Policymakers are encouraged to create regulations that prioritise the interests and well-being of sex workers, including legislation that requires tools to distinguish between voluntary sex work and sex trafficking. This differentiation would be immensely beneficial for both anti-trafficking users and developers, helping prevent unintended negative consequences for this vulnerable group.

Policymakers should also enact laws that address data protection, specifically in the context of anti-trafficking tools. Such legislation would clarify the handling of personal and sensitive data, providing the necessary guidance and safeguards for both users and developers in the anti-trafficking space.

Furthermore, it is essential to establish laws that indemnify victims from the criminal activities they were compelled to engage in as part of their exploitation.

These legal measures would encourage victims to seek assistance and support when anti-trafficking tools provide them with avenues to do so.

## 3.6 Conclusion

This study highlighted the possible impacts of Victim/Trafficker Identification tools and generated guidelines to mitigate any unintended negative consequences associated with these tools.

Understanding the negative consequences and challenges provided valuable information on the opportunities to refine the design and development processes of these tools to ensure responsible outcomes. The recommendations and framework presented in this chapter represent a significant step toward establishing more formal measures that anti-trafficking stakeholders can adopt when developing and utilising anti-trafficking systems, ultimately helping to prevent negative consequences.

## 3.7 Direction of the thesis

As previously mentioned, the initial plan was to focus on the US and these plans were discussed with the participants at the end of the interviews. A few participants, along with other experts consulted during this period, highlighted several concerns about the proposed approach and offered recommendations. The two key suggestions that significantly shape the direction of this thesis are discussed below:

1. Change of geographical focus - The participants and experts highlighted several issues associated with research in the US. Firstly, they discussed the significant risks linked to anti-trafficking tools in the US. Secondly, they emphasised the difficulties in establishing collaborations and acquiring data from online platforms in the US. They suggested that other regions, such as Africa, could be more open to data sharing due to fewer legal restrictions, limited research in this area, and the urgent need to leverage technology in combating human trafficking. However, despite fewer restrictions, researchers must still ensure ethical use of data and minimise any potential harm associated with its

use. Thirdly, they noted a significant gap in research and tools addressing technology-facilitated trafficking in the Global South, particularly in Africa, where the impact of such work could be much greater.

Reflecting on these insights, the decision was made to shift the focus of the study to Africa. This change presented an opportunity to explore an under-researched geographical area, contribute to a deeper understanding of the issue in this region, and lay the groundwork for developing technological interventions. The goal was to maximise the impact of the findings. Furthermore, initial attempts to access recruitment data in the US proved challenging, further supporting the decision to refocus on Africa.

2. The importance of contextual knowledge - The participants also emphasised the critical role of engaging field experts and acquiring contextual knowledge when using technology to detect instances of human trafficking.

   Given the limited literature and the emphasis on stakeholder engagement, it was decided to collaborate with an organisation that could offer insights into human trafficking recruitment in Africa and provide relevant data, thereby enhancing the understanding of the issue's context in the region. The following chapter presents an analysis of datasets on human trafficking recruitment in Africa.

Therefore, with the change in geographical location, this thesis aims to investigate and detect human trafficking recruitment on online platforms in Africa while adopting a responsible innovation approach. Taking a responsible innovation approach means conducting research in a way that minimises potential harm. This is achieved by integrating the principles of the AREA framework into the research process, either directly or primarily through the P3D Framework.

While the P3D Framework is designed for stakeholders developing anti-trafficking tools, this thesis incorporates key recommendations from the framework, particularly in the development of a dataset and the benchmarking of various artificial intelligence approaches on the dataset. Furthermore, the thesis actively engages stakeholders

and integrates expert recommendations, as demonstrated in this chapter, following the Act component of the AREA framework. These considerations ensure that ethical principles and responsible research practices are embedded throughout the process. A summary of how the P3D Framework and RRI influenced this research will be presented in Section 7.

# 4

# Analysis of Human Trafficking Recruitment in Africa

## Contents


**4.1 Introduction . . . . . 108**
**4.2 Background . . . . . 109**
4.2.1 Human Trafficking in Africa . . . . . 109
4.2.2 Human Trafficking Recruitment . . . . . 110
4.2.3 Factors that Contribute to Human Trafficking Recruitment in Africa . . . . . 111
**4.3 Organisation A . . . . . 113**
4.3.1 Background . . . . . 113
**4.4 Methodology . . . . . 118**
4.4.1 Data Collection and Preparation . . . . . 118
4.4.2 Data Visualisation . . . . . 120
**4.5 Results from Analysis of Dataset: Form 1 . . . . . 121**
4.5.1 Overview Analysis: Form 1 . . . . . 121
4.5.2 General Hypothesis Testing . . . . . 128
**4.6 Results from Analysis of Dataset: Form 2 . . . . . 132**
4.6.1 Overview Analysis: Form 2 . . . . . 132
4.6.2 Analysis by Country: Form 2 . . . . . 135
4.6.3 General Hypothesis Testing . . . . . 147
**4.7 Discussion . . . . . 149**
4.7.1 Dataset 1 . . . . . 149
4.7.2 Dataset 2 . . . . . 152
4.7.3 Design Implications . . . . . 153
4.7.4 Direction of Thesis . . . . . 156
**4.8 Conclusion . . . . . 157**

## 4.1 Introduction

The chapter explores the second research question: *How do human traffickers recruit victims both offline and online, and what key factors and patterns contribute to this process?* By addressing this research question, this chapter aims to gain a deeper understanding of human trafficking recruitment in Africa.

A collaboration was established with an organisation that has access to data on human trafficking recruitment in Africa. This chapter refers to the organisation as 'Organisation A' due to their request for anonymity. Organisation A's primary objective is to combat human trafficking recruitment through various initiatives in several regions, including Africa. They provided two datasets, which were analysed using data visualisation techniques, focusing on their operations in 18 Sub-Saharan African countries.

This chapter presents a review of the existing literature and the findings from the analysis of these two datasets. The key contributions of this chapter are as follows:

- Provide insights into human trafficking recruitment in Africa, contributing to knowledge in this underexplored area both offline and online.
- Offer insights that validate various relationships in human trafficking recruitment through hypothesis testing. These hypotheses focus on fraudulent job advertisements on online platforms, aligning with the primary objective of this thesis. They are informed by the data analysis conducted in this chapter and supported by existing literature. However, as most research on fraudulent job postings centres on the Global North, this study explores whether those insights are also applicable to the African context.
- Present a set of design implications, derived from dataset analysis, to support the development of technologies that combat human trafficking recruitment. These implications are crucial as they guide detection systems and shape

policies, ensuring more effective, data-driven prevention and intervention strategies.

- Provide insights, such as indicators of fraudulent job postings, to support the development of datasets in Chapter 6 for benchmarking various AI approaches. These insights ensure that the work accounts for the unique socio-economic, cultural, and technological factors shaping recruitment strategies in Africa, helping machine learning models remain contextually relevant.

This chapter is organised as follows. Section 4.2 provides background information on human trafficking in Africa, with a focus on the recruitment process. Section 4.3 offers background on Organisation A and the origins of the data in the datasets, while Section 4.4 details the methods used to analyse the datasets. Sections 4.5 and 4.6 present insights obtained from the analysis of the datasets. Section 4.7 discusses the findings, their implications and plans for further research. Section 4.8 discusses the limitations of this work and concludes the study.

## 4.2 Background

Africa is home to more than 1.4 billion people (Galal, 2024). It is divided into five main regions (Central Africa, East Africa, Southern Africa, West Africa, and North Africa), comprising 54 countries (Boye, 2022). Sub-Saharan African countries are located south of the Sahara Desert, covering countries in Central Africa, East Africa, Southern Africa, and West Africa. This section discusses human trafficking, focussing on recruitment in Sub-Saharan African countries, where Organisation A operates.

### 4.2.1 Human Trafficking in Africa

Human traffickers are subjecting millions of people in the continent to a variety of exploitative practices, such as forced labour, sexual exploitation, and debt bondage (Allais, 2006; Obokata, 2019; Fabbri et al., 2023). This criminal enterprise is a highly lucrative business for organised criminal networks, facilitated by their ability to transport victims across international borders and between countries (Allais, 2006).

The profitability of human trafficking is perpetuated by several factors, such as low prosecution rates, inadequate legislation, and widespread corruption (Allais, 2006).

Various stakeholders, including governments and non-governmental organisations (NGOs), are actively working to address this issue. Regional bodies such as the African Union (AU) and the Southern African Development Community (SADC) play a pivotal role in the fight against human trafficking. The AU and SADC, composed of 55 and 16 member states, respectively, prioritise preventive measures to address the root causes of trafficking, while governments often emphasise prosecution in their anti-trafficking efforts (Britton and Dean, 2014).

Governments draw upon international frameworks such as the Palermo Protocol and the Trafficking Victims Protection Act (TVPA) to guide their actions. The US Department of State evaluates countries based on their compliance with the minimum standards outlined in the TVPA, assigning scores that reflect the effectiveness of government efforts in combating human trafficking (Britton and Dean, 2014). These scores have substantial implications, the main one being that they can impact whether countries receives funding from the US.

### 4.2.2 Human Trafficking Recruitment

African countries are characterised as the source, transit and destination of human trafficking. The term "source" signifies that these countries are the origins from which trafficking victims are recruited. According to UNICEF, there are primarily two methods through which trafficking victims are recruited in Africa (UNICEF, 2003). The first method involves traffickers approaching victims directly or through their family members. Here, traffickers can manipulate family members, including parents, into handing over the victims as a means to alleviate their financial difficulties (UNICEF, 2003). The second method involves victims or their families approaching traffickers due to difficult circumstances and are desperate for help. In both of these methods, traffickers lure victims with monetary incentives or assurances of an improved lifestyle (Britton and Dean, 2014).

Women are particularly vulnerable to trafficking in Africa, primarily due to gender inequalities and cultural norms (UNICEF, 2003; Britton and Dean, 2014. This vulnerability results in a significant proportion of trafficking victims in Sub-Saharan Africa being women and girls, making them particularly susceptible to various forms of exploitation, including forced prostitution, bride trafficking, and child prostitution (Allais, 2006; Obokata, 2019).

Women also play a significant role in trafficking victims in this region. Research indicates that half of all traffickers in Africa are women (Obokata, 2019). There is a need for a gender inclusive approach to address human trafficking in Africa. These approaches require experts to determine whether women and men use different tactics to recruit victims.

The following section discusses some of the main factors that contribute to the recruitment of human traffickers in Africa.

### 4.2.3 Factors that Contribute to Human Trafficking Recruitment in Africa

#### Unemployment and Poverty

Reports indicate that roughly half of Africa's population is under 25 years of age. However, this young population faces significant challenges in terms of employment. An estimated 12 million young people enter the job market each year, while only 3 million jobs become available (Children International, 2023). This disparity highlights the magnitude of the unemployment problem across the continent, placing millions of young people at risk of various socioeconomic challenges.

In 2023, the unemployment rate in Sub-Saharan Africa was 5.97%, with a labour force participation rate of 43.76% (Macrotrends, 2025). These high unemployment rates, particularly in economically disadvantaged countries within Sub-Saharan Africa, create a conducive environment for traffickers to recruit victims. High unemployment rates contribute to poverty on the continent. Poverty has persisted in many African countries for decades, with limited progress (Onuoha, 2011). A 2021 report indicated that approximately 470 million people in Africa lived below

the poverty threshold of $1.90 per day (UN Trade & Development, 2021). These economic challenges facing African nations have contributed significantly to the high number of victims of trafficking on the continent (UNICEF, 2003).

### Education

UNESCO data reveal that Sub-Saharan Africa is grappling with the highest levels of education exclusion, with more than 60% of young people aged 15-16 years of age not attending school (UNESCO, n.d.). The exclusion rate from education is significantly higher for adolescent girls (36%) than for boys (32%).

Furthermore, education systems in many African countries show a variety of issues, including limited resources, a lack of teachers, and inadequate infrastructure (UNESCO, n.d.). These challenges significantly hinder students' access to quality education. Insufficient resources and teacher shortages, in particular, can result in unsupervised students, making them more vulnerable to recruitment by traffickers. In addition, a lack of education leads to high unemployment rates, which further contributes to individuals being recruited into human trafficking.

### Political Instability and Corruption

Political instability and corruption often play an important role in exacerbating the problem of human trafficking in Africa (Fabbri et al., 2023; Fedele, 2021; Global Slavery Index, 2018). A few studies (Fabbri et al., 2023; Fedele, 2021; Global Slavery Index, 2018) state that the high incidence of trafficking in Africa is the result of governance-related issues. These include ineffective government leadership, political instability, and limited government capacity to combat trafficking effectively.

There are several concerns about the way African governments operate in fighting human trafficking. The following are the identified concerns (Onuoha, 2011):

- African leaders often prioritise their interests over the allocation of resources and efforts to tackle human trafficking, with a greater concern for electoral politics.

- There is a lack of effective policies and strategies to combat human trafficking within the region.
- Collaboration between African governments and non-governmental organisations (NGOs) in the fight against human trafficking is often lacking or ineffective.
- Corruption within law enforcement agencies creates an environment in which traffickers can easily evade prosecution.
- The scarcity of resources to address societal needs creates an environment conducive to traffickers' recruitment efforts.

These multifaceted challenges have contributed to the persistence of human trafficking on the continent. Addressing this issue requires multifaceted solutions to mitigate these challenges and other related problems. Furthermore, more research is needed to better understand the profiles of victims, with particular emphasis on factors such as educational level, occupational status, and other variables that can shed light on the circumstances that led to their recruitment.

## 4.3 Organisation A

This section provides further details on the work of Organisation A and its methods of data collection, which form the datasets analysed in this chapter.

### 4.3.1 Background

Organisation A employs diverse methodologies in the 18 African countries where it operates. The predominant approach involves positioning Organisation A personnel at border control points, such as arrival and departure areas, to identify individuals potentially at risk of recruitment into human trafficking. Organisation A collaborates with immigration authorities and law enforcement agencies to interrogate potential victims and traffickers when trafficking is suspected. In countries where formal partnerships with government bodies have not been established, Organisation A

conducts alternative initiatives, investigating both digital and physical environments to identify potential victims. Regardless of the specific strategy employed, Organisation A's primary objective is to prevent recruitment before or during the transportation phase, thereby safeguarding individuals from traumatic and exploitative circumstances such as forced labour and sexual exploitation.

This chapter focusses primarily on the operations of Organisation A at the border control points, a methodology used in all but one of the African countries. In that exception, the emphasis is on analysing recruitment through digital platforms. Chapter 5 provides further details on the country that focusses on digital platforms. The core processes employed by Organisation A at the border control points are summarised in Figure 4.1.

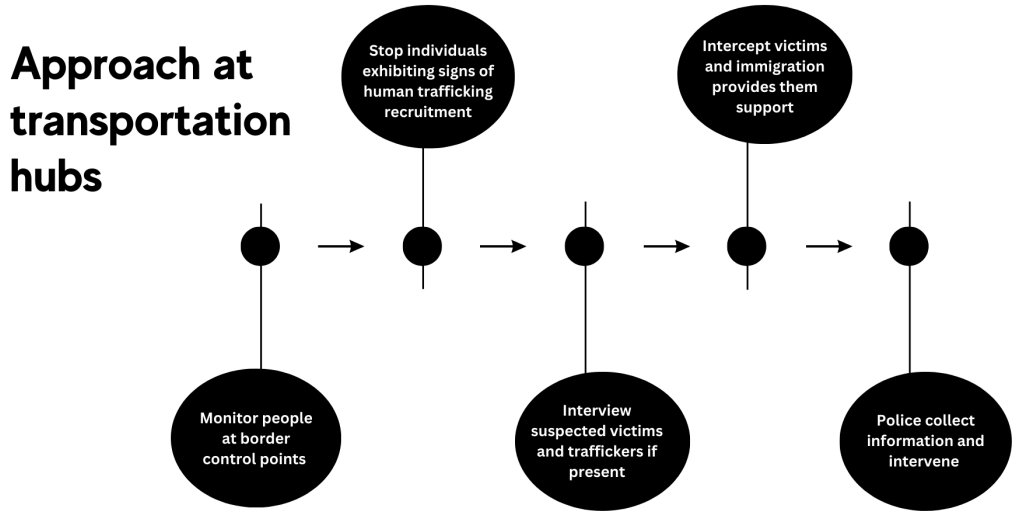


**Figure 4.1:** Organisation A Process at Border Control Points

The process begins with the surveillance of individuals at the border points and the stopping of those who display indicators of human trafficking. Organisation A concentrates on individuals from high-risk countries, observing signs such as unaccompanied minors or suspicious behaviour. Additional indicators are discussed in Sections 4.5 and 4.6. If there are sufficient grounds for suspicion, Organisation A

intervenes by stopping either potential victims travelling alone or both the victims and suspected traffickers when they are travelling together.

Upon stopping suspected individuals, the staff conduct interviews to gather additional information and confirm whether potential victims are at risk of human trafficking. If they determine that potential victims are at high risk, Organisation A intercepts them, preventing the continuation of their journey. In this chapter, "intercept" refers to Organisation A's successful intervention in preventing a victim from being recruited. After interception, Organisation A provides the information to law enforcement authorities, who subsequently assume responsibility for the case.

When Organisation A successfully prevents potential victims from being recruited, its personnel collects detailed information about the recruitment process directly from victims. They also document the factors that led to the interception, such as suspicious behaviour or trafficking patterns. This information is recorded in spreadsheets and constitutes the datasets analysed in this chapter. Organisation A uses two distinct forms to collect this information after intercepting the suspected victims. These forms will be referred to as Forms 1 and 2 in this chapter.

The following section explores the specific details of these forms, explaining their purposes, and the data they are designed to capture. However, due to privacy and anonymity concerns, the forms themselves will not be included in this thesis.

### Form 1

A part of the interview process involves determining whether the case presents a high risk of human trafficking or whether there is sufficient evidence to suggest human trafficking. Form 1 records information detailing why Organisation A staff consider a case to be high risk or have evidence of human trafficking.

To determine whether there is a high risk of human trafficking, the staff uses three criteria as described in Form 1. These criteria are described below:

- The first criterion involves assessing whether suspected victims fit the profile that traffickers often target. Examples of popular profiles used in such cases

include being a runaway, recent enslavement, unemployment, and living in impoverished conditions.

- The second criterion involves evaluating whether suspected victims are moving to areas known for trafficking activities. Examples include destinations associated with domestic work. These areas may expose individuals to forced labour, exploitation, or other forms of human trafficking.

- The final criterion involves determining whether suspected victims lack the resources necessary to ensure their safety. Officials pay attention to issues such as people who cannot communicate in the language of the destination country and unaccompanied minors who lack a guardian.

Form 1 provides different options for each criterion, allowing the staff to select and record the relevant options corresponding to each intercepted individual.

For a case to be considered to have **evidence of human trafficking**, it must show signs of "signs of control". These signs of control help officials identify instances of manipulation, coercion, or deceit used to exploit a victim. The form contains predefined signs of control, each assigned a specific score. These scores vary for each sign, creating an assessment framework.

Each country establishes a minimum numerical threshold that a case must reach to be considered evidence of human trafficking based on cumulative scores. If the total calculated score of a case meets or exceeds this threshold, the case is acknowledged as having evidence of human trafficking. This scoring approach ensures that only cases that meet the predefined criteria are classified as evidence of human trafficking, providing a consistent evaluation process. Cases that do not meet the minimum threshold but satisfy all three criteria of high-risk indicators, as described above, are recorded as "high risk of human trafficking". Organisation A always intervenes in both instances where cases are classified as either "high risk of human trafficking" or "evidence of human trafficking".

Organisation A also uses Form 1 to record personal information about potential victims and suspected traffickers if present.

### Form 2

Form 2 is used when there is evidence of human trafficking based on the approach described above. Ideally, all cases marked as evidence of trafficking on Form 1 should be systematically documented on Form 2. However, only a few countries consistently record on Form 2. As a result, the dataset based on Form 2 exclusively represents the countries that adhere to this documentation process.

Information stored in Form 2 includes:

- Details about potential victims, including their age, gender, and family-related information.
- Information about the recruitment process, such as the method used for recruitment, the amount of time potential victims were in contact with traffickers, and the financial incentives offered for job opportunities when fraudulent means were involved.
- Information about previous travel details for potential victims before arriving at the border. Data on this include the countries and cities travelled by the victim, travel documents, and the specific routes taken.
- Information on the legal aspects of the case, such as whether legal actions were initiated against the trafficker, the reasons for not proceeding with the case if it was not filed, details of the interactions between law enforcement and potential victims, dates of visits to the police station, and information regarding the victim's statement.
- Personal information about all individuals involved in the case, including details of the locations to which they travelled before interception and information about the vehicles they used.

The datasets analysed in this chapter are based on information extracted from these two forms. The following section provides insights into the nature of these datasets and the methodology used for their analysis.

## 4.4 Methodology

This section discusses the data collection, preparation, and analysis process.

### 4.4.1 Data Collection and Preparation

The initial contact with Organisation A was made by email after obtaining their contact details from their website. In this email, I introduced my project and requested a potential collaboration. Subsequently, the administrator facilitated a connection between me and the head of data science, who later agreed to a collaboration and sharing of the datasets. The period to identify Organisation A and gain access to the datasets spanned a period of five months.

Organisation A provided two datasets containing data from the last five years. It is crucial to note that the organisation anonymised both datasets before sharing the information. This helped ensure that any personal identifiable information was removed before accessing the data.

The data were securely stored on a OneDrive service linked to my Nexus account and also on my local machine. Both the OneDrive service and the local machine were password protected, requiring permission to access. The storage space on OneDrive does not automatically backup files, and there is a deletion provision if it is unused for more than 180 days. The following sections discuss the two datasets in detail.

#### Dataset 1

Dataset 1 is derived from Form 1 and includes fields that capture information on the state of victims and suspects during interceptions. Each entry represents a potential victim who may be involved in a trafficking situation. When multiple victims were intercepted within a single case, the organisation still records each victim as a separate occurrence, but the victims are connected with a unique identification number.

The dataset provides information on whether a case is classified as high risk for human trafficking or meets the minimum threshold for being considered as evidence

of human trafficking. It also documents the specific criteria and indicators used for classifying a case as evidence of human trafficking.

Dataset 1 contains qualitative and quantitative data and consists of 18,146 rows and 83 columns. Before analysing the data, it was cleaned and preprocessed to enhance its quality and prepare it for analysis. The following main steps were taken to clean and preprocess the dataset:

- Correcting spelling mistakes and standardising spelling variations.
- Reformatting the date and time columns into multiple columns representing day, time, and month.
- Handling null values appropriately.
- Removing extra spaces and non-textual characters to properly format text and prevent errors during analysis.
- Formatting data to ensure uniform casing.
- Resolving variations in the data, such as mapping similar city names to corresponding country names to maintain consistency in broader column details.

Appendix B.1.1 provides more details on these cleaning and preprocessing steps.

**Dataset 2**

Dataset 2 originates from Form 2 and includes various fields designed to record information about victims, the recruitment process, the motivations behind the recruitment, the nature of exploitation, the legal aspects of the case, and details of travel and destination.

The second dataset also has qualitative and quantitative data and consists of 2446 rows and 82 columns. The significance of this dataset lies in its ability to provide information on critical aspects of recruitment. These insights are valuable in formulating strategies to combat human trafficking recruitment. Dataset 2 also

underwent various cleaning and preprocessing steps similar to those described above for Dataset 1. However, additional steps were taken for Dataset 2, which included:

- Removing columns that were not relevant to addressing the research questions in this sub-study.
- Merging of certain columns to enhance analysis and facilitate pattern identification.
- Replacement of outliers with the mode of the respective field to maintain data consistency.

Additional details regarding these cleaning and preprocessing steps are provided in the Appendix B.1.2.

### 4.4.2 Data Visualisation

Data visualisation (Aparicio et al., 2014) served as the primary method to analyse both datasets. Data visualisation is an interdisciplinary method that draws input from psychology, computer science, data mining techniques, and graphic design. It is a critical step in data science (Aparicio et al., 2014). Various popular techniques, such as box plots, histograms, heat maps, charts, tree maps, and word cloud/network diagrams, contribute to its versatility.

Data visualisation was chosen for its ability to provide valuable information for decision-making, identify areas for future research, and uncover patterns within the data. It allows for a visual representation of the data, making it easy to identify patterns, outliers, and trends through graphs and charts. With an increasing volume of data, visualisation becomes crucial, as it enables individuals to grasp patterns more easily through the use of colours and figures, facilitating effective communication within the academic and research community (Great Learning, 2024; Hashemi-Pour, 2024). Through data visualisation, researchers can pinpoint areas that impact their research phenomenon and identify avenues for improvement. Furthermore, visualisation improves the quality of the decision-making process (Eberhard, 2023).

This study used Python for visualisation due to its extensive libraries and coding efficiency (Yada, 2022). The main libraries employed for data visualisation include:

- Pandas - It was used mainly to clean and organise the data before visualisation (Pandas, 2024).
- Panel - It was used to create a dynamic dashboard to display graphs, chosen for its compatibility with Jupyter Notebooks, the primary coding environment used in this study (Panel, 2024).
- Holoviews - It was used to create interactive plots, offering controls for graph manipulation, such as accessible data saving and zooming capabilities (HoloViews, 2024).
- Matplotlib - This helped generate 2D static plots to visualise various data and also served as a backup when encountering issues with Holoviews, providing ease of multi-graph manipulation (Matplotlib, 2024).
- Geopandas and Follium - It helped to visualise geospatial data, particularly because it was from different African countries. Folium played a crucial role in enhancing interactivity in geospatial representations (GeoPandas, 2024).

Different visual representations were selected on the basis of the data being analysed. Bar charts, for example, were predominantly used for categorical data.

## 4.5 Results from Analysis of Dataset: Form 1

This section discusses the insights gained from data visualisation and hypothesis testing from the first dataset.

### 4.5.1 Overview Analysis: Form 1

The data show that there is a discernible increase in intercepted victims over time, as shown in Figure 4.2. This increase could signify that the anti-trafficking strategies implemented by Organisation A are becoming more efficient in identifying potential

victims. However, this increase may also suggest that traffickers are successfully reaching more victims in Africa than in previous years. The analysis also reveals that the number of traffickers involved in trafficking cases is increasing during the same period, which could indicate an expansion in trafficking networks.

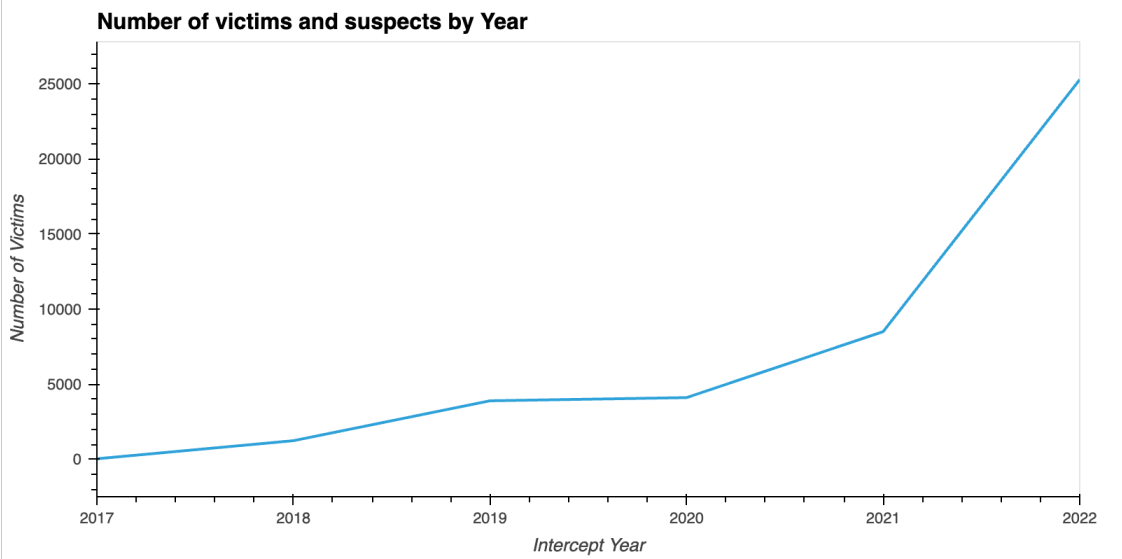


**Figure 4.2:** Total Victims Over the Years

Furthermore, the data show a positive correlation between the number of victims and years, with a particularly significant increase observed in 2022. This aligns with the literature indicating a global increase in human trafficking victims in 2022, driven by various factors, such as wars and conflicts (United Nations: Office on Drugs and Crime, 2023; Statista Research Department, 2024; The World Bank, 2022). Previous studies indicate a decrease in victims detected in 2020, including a 12% decline in Sub-Saharan Africa (United Nations: Office on Drugs and Crime, 2023; Statista Research Department, 2024), while Organisation A reported a 5% increase in the number of victims detected. The global decline in human trafficking cases has been attributed to various factors, such as the COVID-19 pandemic, which affected trafficking flows across borders and restricted access to locations that facilitate trafficking activities, especially in low and middle income countries (United Nations: Office on Drugs and Crime, 2023). The increase in cases for Organisation A in 2020 is likely attributable to other factors such as improved intervention strategies.

The victims intercepted in this dataset represent a diverse range of nationalities, with significant numbers coming from countries such as Malawi, Angola, Zimbabwe, Uganda, and Mozambique (see Fig. 4.3). Normalising the number of trafficking victims by population provides a more accurate representation of the prevalence of human trafficking. After normalising the data by population, the countries with the highest number of victims relative to their population include Namibia, Malawi, Zimbabwe, Lesotho, and Liberia.

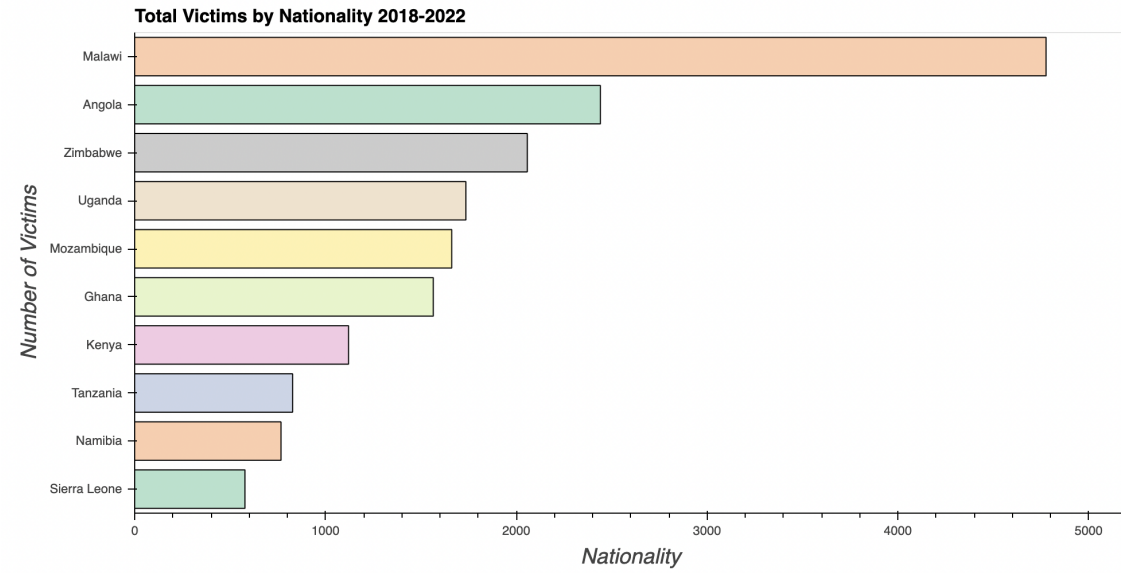


**Figure 4.3:** Total Victims- Top 10 Nationalities

Not all victims identified in this dataset originated in Africa; some were recruited from regions outside Africa and subsequently brought to African countries. This highlights the global nature of human trafficking networks and reflects a common phenomenon in which victims are transported and detected in regions outside their origin. However, most of the individuals intercepted were African. Among the victims recruited outside of Africa, the majority came from Asia.

The dataset also sheds light on the predominant roles that staff members play, in contrast to law enforcement, in intercepting victims, as visually represented in Figure 4.4. There were notably few cases in which the **victims approached the staff independently**.

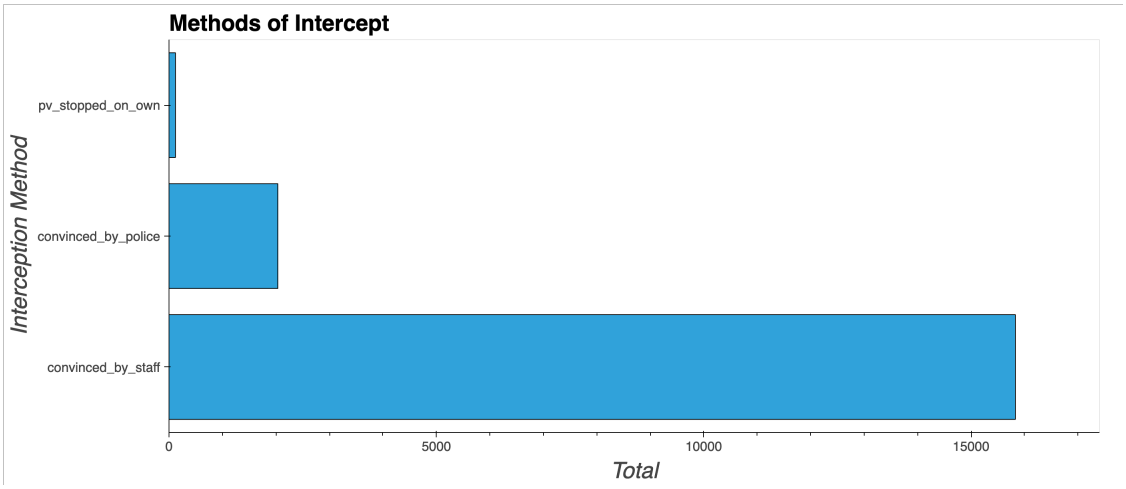


**Figure 4.4:** Methods of Intercepting Victims

As discussed previously, the organisation employs three criteria to determine whether there is a high risk of trafficking: 1) profile, 2) industry, and 3) signs of vulnerability.

The data highlight that the most prevalent profiles among the intercepted victims were of children and individuals with a youthful appearance, many of whom were runaway or originating from impoverished backgrounds. Migrants were also frequently intercepted.

Regarding the industries associated with potential trafficking situations, domestic work emerged as the most common, followed closely by farm work, dance bars, and the hotel industry. Domestic work constitutes a significant industry in Sub-Saharan Africa, with an estimated 9.6 million domestic workers (Nayupe et al., 2023).

Figure 4.5 provides an analysis of the common indicators of vulnerability observed among victims. A significant proportion of victims lacked adequate resources, such as money to live in the country or access to essentials such as a phone, at the time of interception. In addition, there were instances where victims had no information about their intended destination or could not communicate in the destination country's language. The dataset also highlights that many of the victims intercepted were **minors without available guardians**, making them particularly vulnerable to exploitation.

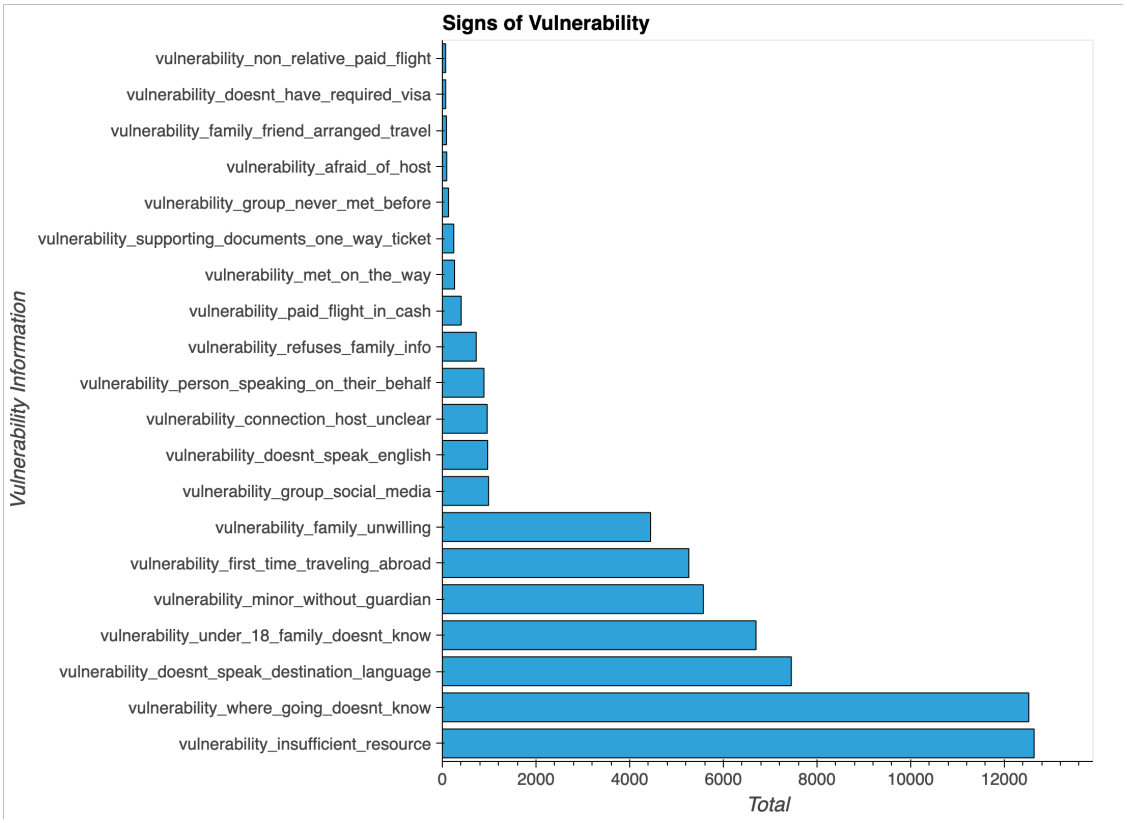


**Figure 4.5:** Signs of Vulnerability

Figure 4.6 illustrates further patterns among high-risk individuals. A prevalent pattern involved people travelling without any luggage for what appeared to be a lengthy journey or carrying substantial luggage while claiming that their stay was of short duration. Additional recurring patterns included individuals who were unable to **provide verifiable details about their employment** or were caught in lies. Some victims openly admitted to deceiving officials.

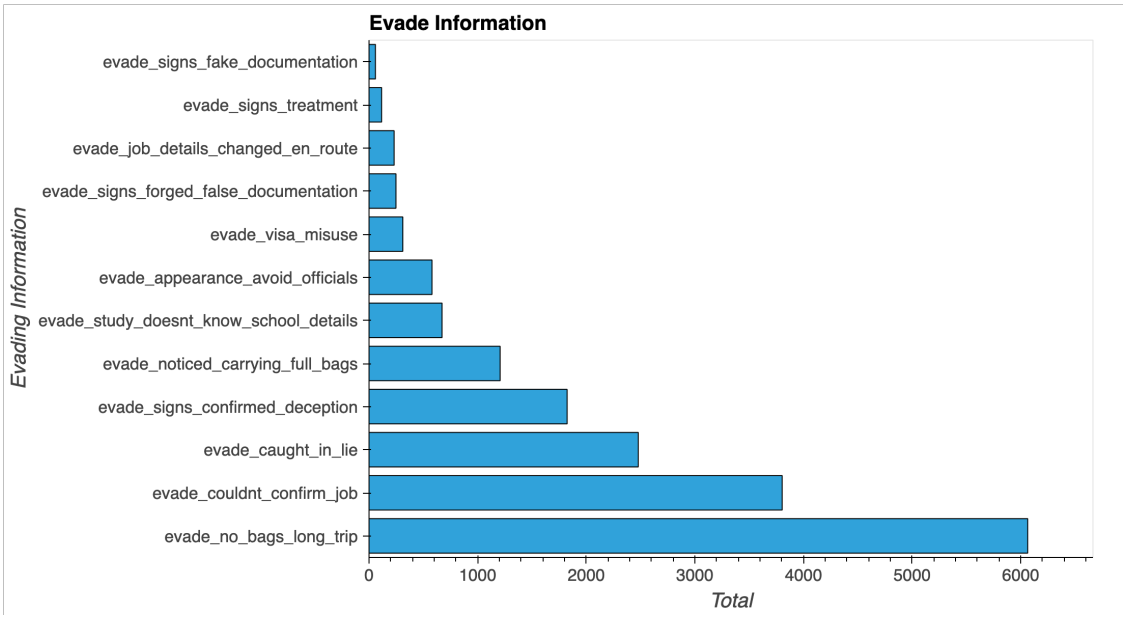


**Figure 4.6:** Methods of Evading Interceptions

As previously mentioned, Organisation A staff typically seek various signs of control when assessing a situation. One of the main ways they identify signs of control is by asking about the reasons a victim leaves their country. Figure 4.7 summarises these reasons. The primary reason for leaving is related to debts, which means that individuals are trafficked into exploitative situations due to their debts. Another common reason is coercion through threats.

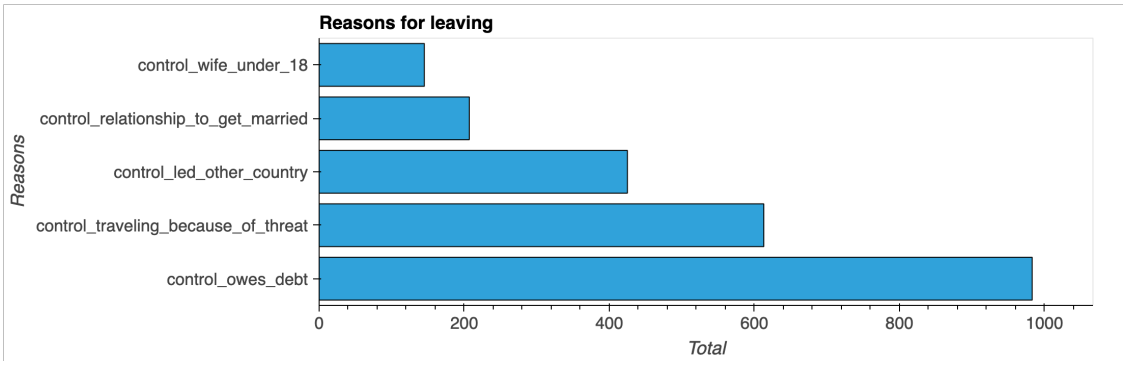


**Figure 4.7:** Control Means - Reasons for Leaving

If a victim leaves due to marriage, the staff is required to enquire about the specific details of the marriage, mainly the duration. Figure 4.8 illustrates that most of the victims intercepted entered marriage within a relatively short period of time of 2 to 8 weeks after meeting each other. A short period of time raises suspicion.

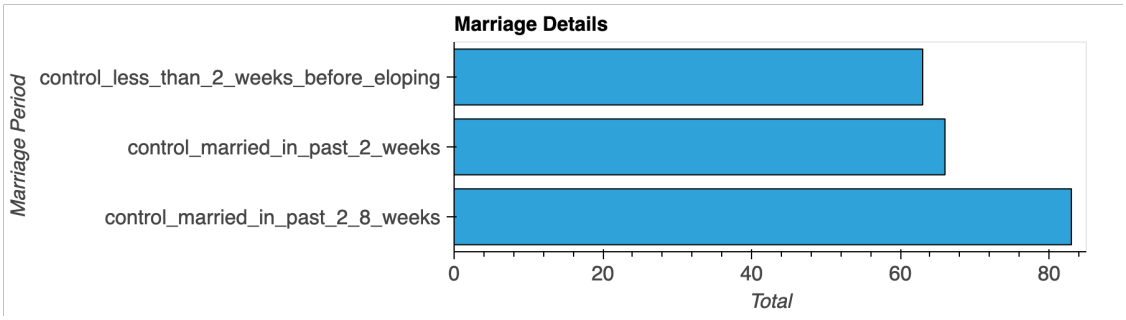


**Figure 4.8:** Control Means - Marriage Details

If the reason for leaving is because of work, the staff ask for the **details of the job** and the qualifications of the victim. In this dataset, the tactics used most frequently by traffickers to trap individuals are the promise of higher paying jobs and the promise of a fake job, as shown in Figure 4.9.

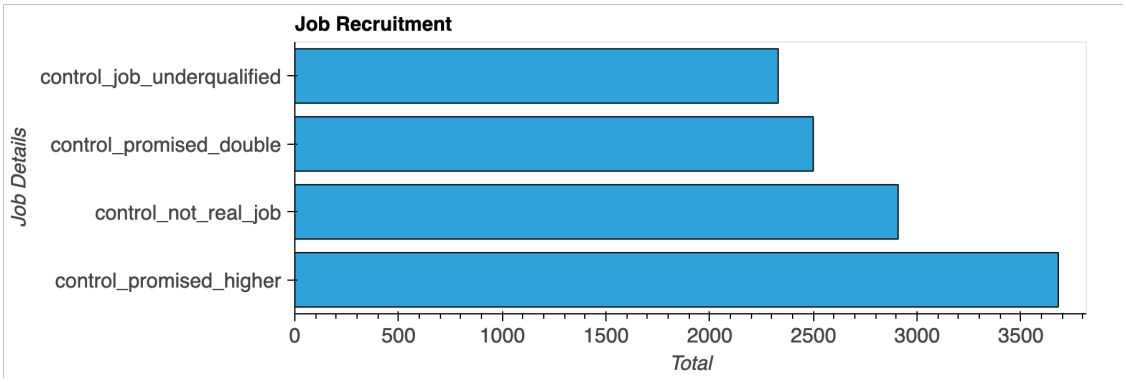


**Figure 4.9:** Control Means - Job Details

The staff also look for signs that suspected traffickers are controlling victims. Figure 4.10 illustrates some of these signs of control noted by Organisation A staff.

The most common vulnerability is that traffickers cover the travel expenses of potential victims. This information can be valuable for identifying possible trafficking situations. Another prevalent control method is the encounter of a minor who has been forcibly separated from their family.

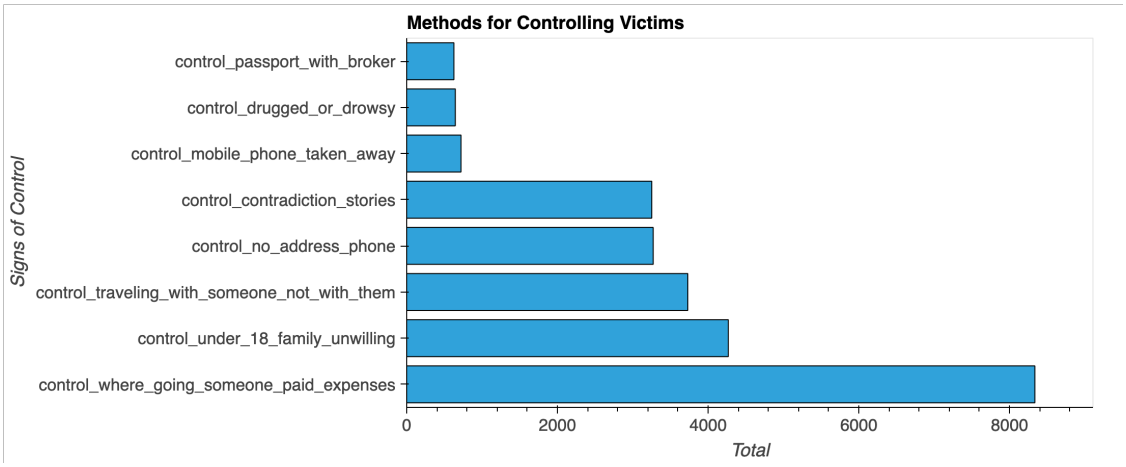


**Figure 4.10:** Control Means - Conditions of Victims

Figure 4.11 summarises the cases classified as high-risk for trafficking and those with evidence of trafficking. Only a few cases were incorrectly intercepted, indicating that the officials responsible for these intercepts are well trained and highly accurate, with an error rate of less than 1%.

This is further supported by additional data collected from officials, who were asked to rate their level of certainty about whether a case involved actual trafficking, using a scale ranging from 1 to 5. Most of the officials assigned a rating of 5, which indicates a high confidence in their intercepts, followed by a rating of 4.

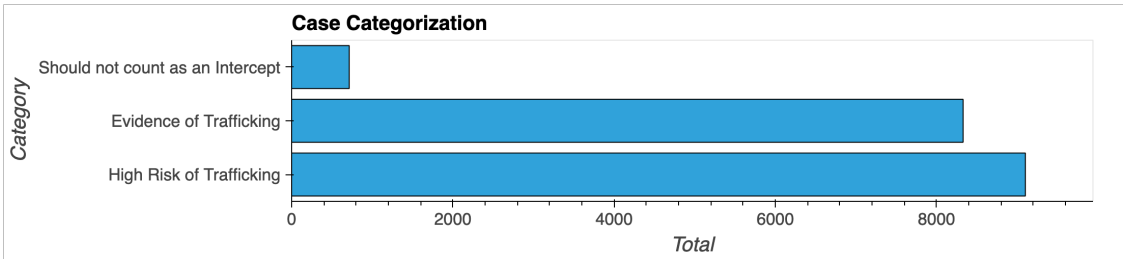


**Figure 4.11:** Summary of Cases

To better understand the data, several hypotheses were tested. The results are described in the following section.

### 4.5.2 General Hypothesis Testing

This section provides a statistical analysis aimed at exploring relationships related to fraudulent job opportunities on online platforms. As mentioned in the Introduction

section of this chapter, all hypotheses in this chapter are based on the analysis of the datasets and a previous review of the literature.

Three hypotheses are tested using the null hypothesis ($H_0$) versus the alternative hypothesis ($H_1$). The null hypothesis states that there is no relationship between the variables and that any observed result is purely due to chance (Vaidya, 2023). The null hypothesis is accepted if there is no significant relationship between the two variables and the alternative hypothesis is accepted if a significant relationship exists. The chosen significance level for this analysis is set at 0.05, which guides the decision on whether to accept or reject the null hypothesis. Specifically, if the p-value is greater than 0.05, the research leans towards accepting the null hypothesis, while a p-value less than 0.05 leads to favouring the alternative hypothesis. The most common p-value threshold, 0.05, is used as it suffices for this study (McLeod, n.d.).

The chi-square test of independence serves as the primary method to test hypotheses because the test mainly involved categorical variables (Vaidya, 2023). The chi-square test is a known method for testing two categorical variables (Vaidya, 2023). Further investigation is done on the analysis using post hoc statistical assessments, specifically residual analysis. Post hoc analyses are used after the initial analysis to investigate the relationships within the data (category values) and identify the sources of differences (Foster et al., 2024). Residual analysis identifies the category values that make the greatest contribution to the results of the chi-square test when there is a significant relationship (Sharpe, 2015).

Before performing the Chi-square test, the suitability of the data was assessed by examining the values in each cell of the contingency table. A contingency table is a statistical tool used to compare the relationship between two categorical variables (Durso, n.d.). It helps determine whether there is a dependency between them by displaying the frequency of observations for each possible combination of category values (Frost, n.d.).

In this analysis, a contingency table was created using the Python crosstab function, which calculates the number of cases for each combination of the two

categorical variables (Pandas, 2025). Each cell in the table represents the count of cases that fall into a specific combination of the two variables being analysed.

For the Chi-square test to be valid, it is recommended that each cell value exceeds five to ensure statistical reliability (Schnell, n.d.; Bobbitt, 2021; JMP, 2023). All statistical tests were performed using Python.

**Are cases involving social media groups more likely to be confirmed as high risk for human trafficking, or to have sufficient evidence to confirm human trafficking?**

This hypothesis is tested because the data analysed, as shown in Figure 4.5, indicates that many of the victims intercepted online were recruited through social media groups. The figure illustrates various vulnerabilities observed among victims, with recruitment through social media groups being the most common vulnerability identified on online platforms.

The aim is to determine whether cases involving recruitment through social media groups are frequently classified as high-risk or have sufficient evidence to indicate human trafficking. As outlined in Sections 4.3.1 and 4.4.1, staff classify each case as either high risk of human trafficking or as having evidence of trafficking. This case classification is recorded in a column labelled **Verified Evidence Category**. To test this hypothesis, the relationship between social media groups and verified evidence category was examined.

$H_0$: There is no relationship between social media groups and the verified evidence category.

$H_1$: There is a relationship between social media groups and the verified evidence category.

Each cell value exceeded five; therefore, the Chi-square test was deemed appropriate to test the hypothesis. Based on the test results, a chi-square value of 653.9 was obtained, with a p-value of 1.0265e-142, and a degree of freedom of two. The results revealed a significant relationship between social media groups and the verified evidence category, leading to the rejection of the null hypothesis.

To further analyse the relationship between social media groups and verified evidence categorisation, additional tests were performed using residual analysis. The results demonstrated a stronger relationship between social media groups and evidence of trafficking, unlike the other options.

**Is there a relationship between jobs that are not real and social media groups?**

The hypothesis tested aimed to determine whether there was a connection between social media groups and non-existent jobs. Figure 4.9 indicated that the staff often discover that a job is not real during their investigations. Therefore, the hypothesis sought to establish whether people are likely to encounter a substantial number of fake job opportunities in social media groups. The relationship between social media groups and non-existent jobs was tested to explore this connection.

$H_0$: There is no relationship between social media groups and non-existent jobs

$H_1$: There is a relationship between social media groups and non-existent jobs

The chi-square test was considered appropriate since each cell value exceeded 5 in the contingency table, supporting the suitability of the test. The results returned a chi-square value of 2723.69, a p-value of 0.0, and a degree of freedom of one. Consequently, the null hypothesis is rejected, indicating a significant relationship between social media groups and jobs that do not exist.

**Can officials easily verify cases of trafficking when the victim is offered a higher salary for a job?**

The hypothesis tested examines whether there is a relationship between the attributes of verified evidence and the salaries offered for jobs. Figure 4.9 indicates that victims intercepted were often promised higher salaries than those typically associated with the jobs. This hypothesis assesses whether most cases involving higher salaries are high-risk or whether there is sufficient evidence of human trafficking. This analysis can help determine how strong an indicator salary might be during investigations.

$H_0$: There is no relationship between the verified evidence category and higher salaries

$H_1$: There is a relationship between the verified evidence category and higher salaries

The chi-square test was considered appropriate since each cell value exceeded 5 in the contingency table, supporting the suitability of the test. The results returned a chi-square value of 758.0, a p-value of 7.2133e-144, and a degree of freedom of two. Consequently, the null hypothesis is rejected, indicating a significant relationship between higher salaries and the verified evidence category.

To further explore the relationship between verified evidence categorisation and higher salaries, a post hoc analysis was performed using residual analysis. The results showed a stronger relationship between higher salaries and evidence of trafficking, indicating that cases involving job offers with higher salaries were confirmed to be associated with human trafficking.

## 4.6 Results from Analysis of Dataset: Form 2

This section summarises the insights derived from Dataset 2.

### 4.6.1 Overview Analysis: Form 2

Image 4.12 shows that the majority of victims in absolute numbers came from Uganda, Malawi, Angola, and South Africa.

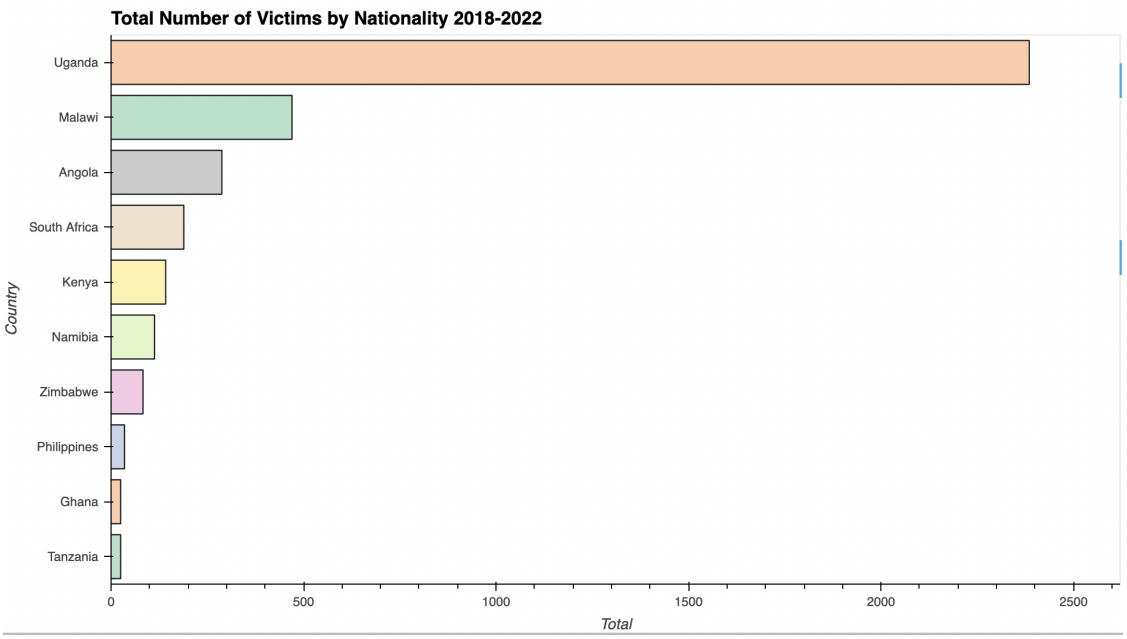


**Figure 4.12:** Total Number of Victims Based on their Nationality

According to the dataset, a significant proportion of victims have **completed secondary school education** or are currently **attending secondary school**. The analysis also revealed that the main recruitment method in the dataset is through **traffickers who directly approach victims**, as shown in Figure 4.13, followed by **recruitment through family and friends**. Analysis of the **five countries** with the highest number of victims revealed that **family and friends** are the main recruitment methods in the countries, followed by **job advertisements**.

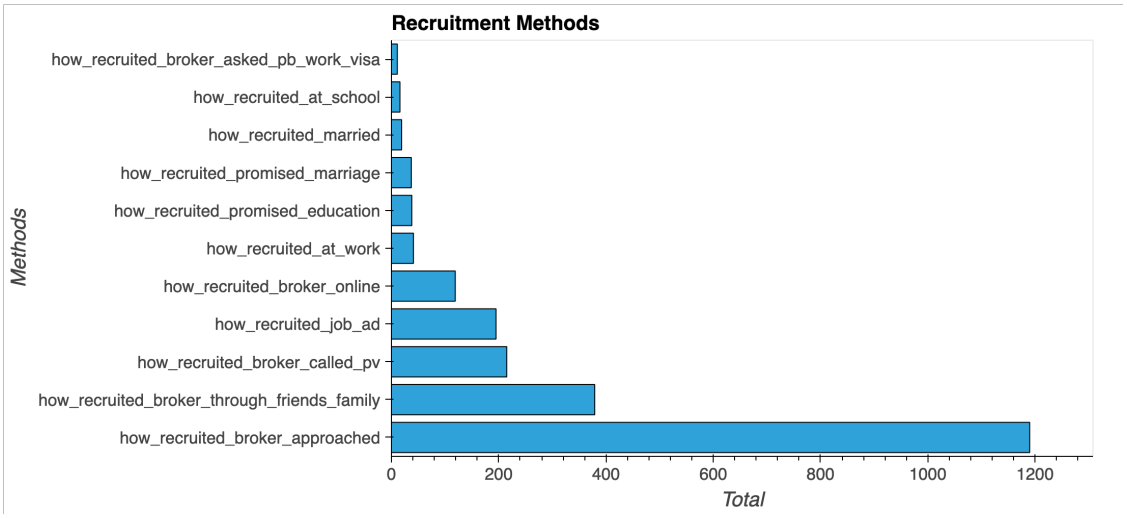


**Figure 4.13:** The General Recruitment Methods

The purpose of recruitment plays a central role in understanding human trafficking. According to the dataset, the most common reason for recruitment is hotel and household jobs, as illustrated in Figure 4.14. This shows that most victims leave for job opportunities. Considering that many victims have attained secondary education or are currently in secondary school, it is plausible that they mainly apply for household and hotel jobs because they typically require minimal educational qualifications.

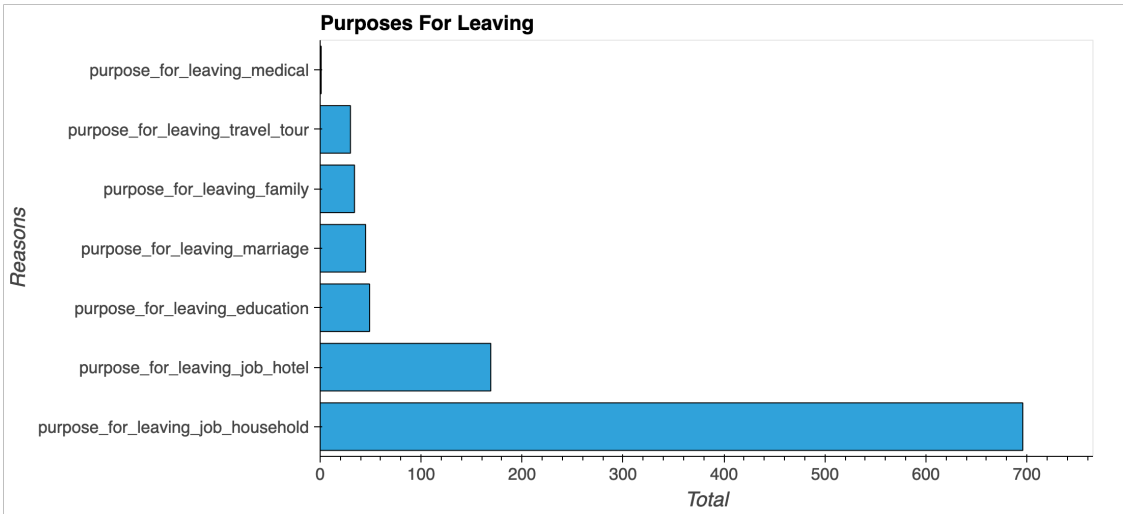


**Figure 4.14:** Methods of Enticing Victims into Trafficking Life

Forced labour is the most prevalent form of exploitation among the victims, followed by sexual abuse and forced prostitution. This aligns with reports by the United Nations that forced labour is the most prevalent form of trafficking in Sub-Saharan Africa (United Nations Office on Drugs and Crime, 2023).

Reports have indicated that African victims are frequently trafficked to neighbouring countries within the region, often from economically disadvantaged nations to those with more advanced economies (United Nations: Office on Drugs and Crime, 2023). Namibia, South Africa, and Kenya were the predominant destination countries for victims in the countries examined. These nations are among the most economically developed in the region, and traffickers are likely to entice victims with the prospect of improved living conditions.

The following section discusses the top five countries with the highest total number of victims to better understand recruitment practices in greater detail.

### 4.6.2 Analysis by Country: Form 2

The section below discusses five countries with the most victims in the dataset: Uganda, Malawi, Angola, Kenya, and South Africa.

#### Uganda

As discussed previously, Uganda has the highest number of victims in the dataset. There was a significant increase in victims in 2022 compared to previous years. Figure 4.15 highlights that most of the victims in Uganda are women and predominantly have a secondary school education.

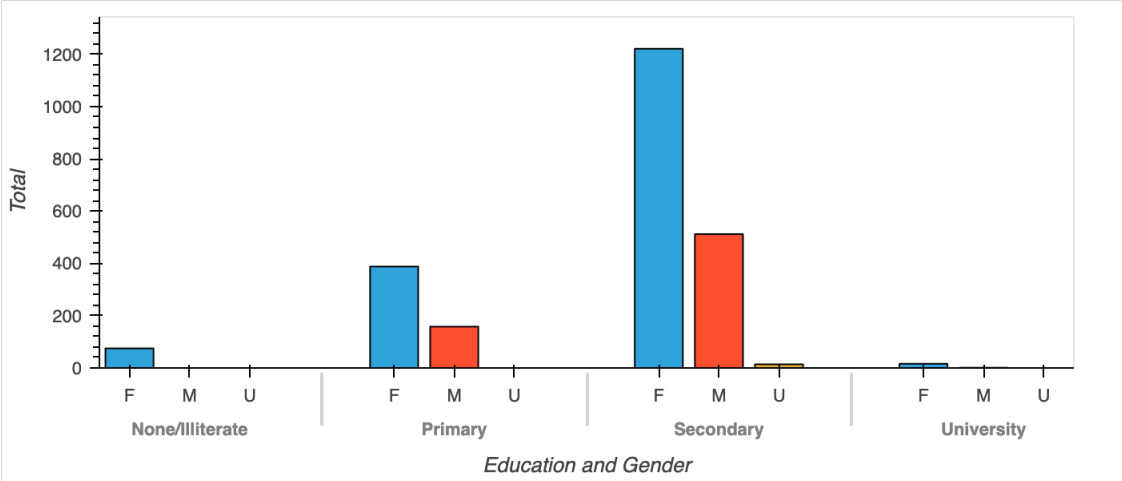


**Figure 4.15:** Uganda Education Statistics

The data presented in Figure 4.16 further reinforce the observation that a significant number of victims in Uganda are current secondary school students. Uganda implemented a programme that offers free secondary school education to students of poor backgrounds who achieve certain grades, thus improving enrolment and the gender gap in schools (Kavuma, 2011). This programme could explain why many victims have secondary education due to high enrolment in secondary schools.

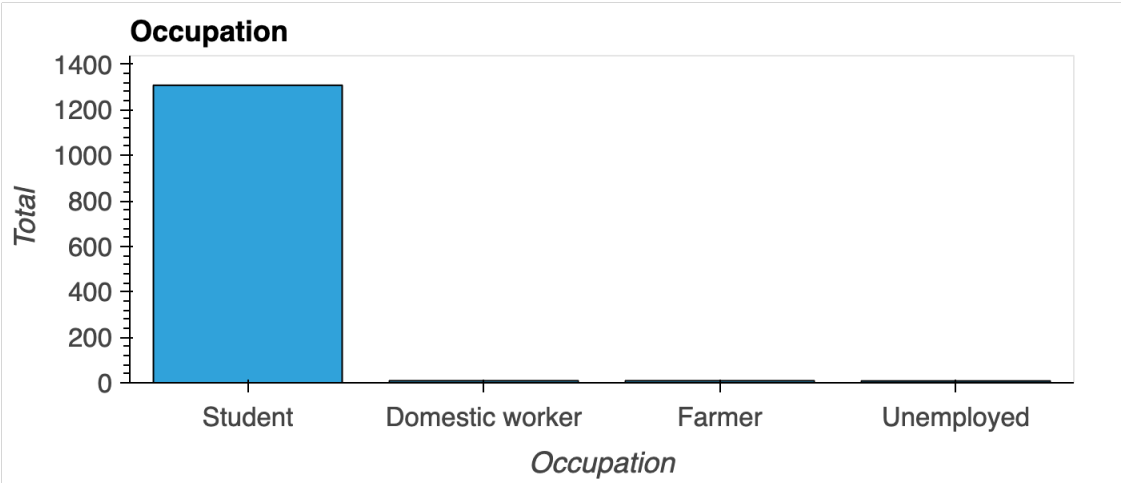


**Figure 4.16:** Uganda Top Occupations

Uganda's most common recruitment method involves victims being approached directly by suspected traffickers, as seen in Figure 4.17. The other most common methods include traffickers who contact potential victims through phone calls and recruit victims through family and friends.

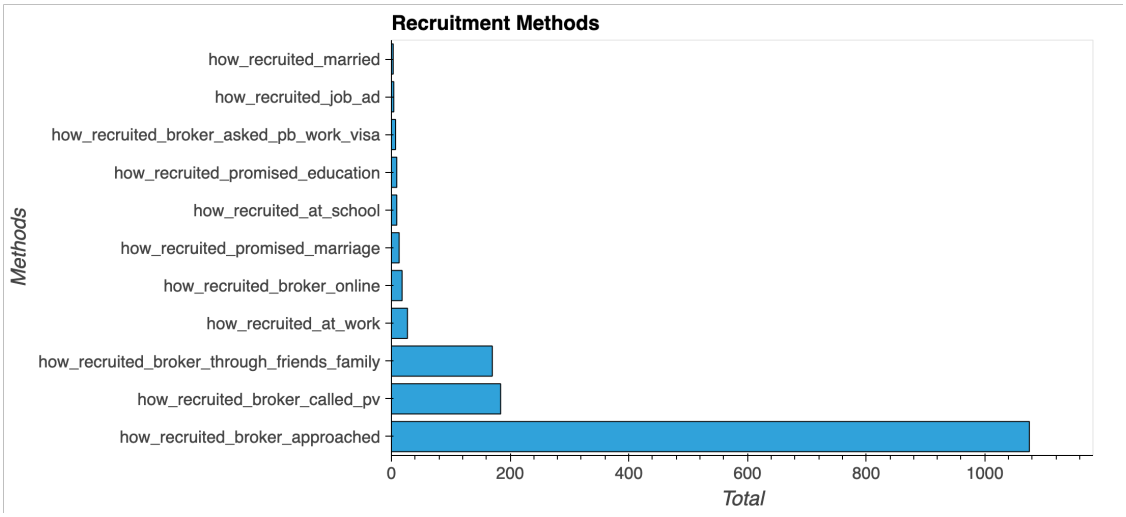


**Figure 4.17:** Uganda Main Recruitment Methods

The primary reason for victims leaving Uganda, as shown in Figure 4.18, was the promise of job opportunities, which is particularly attractive to secondary school students from poor backgrounds who seek to improve their economic status. According to a previous study (Ananny and Crawford, 2018), secondary school students in Uganda are vulnerable to high dropout rates. The high dropout

rate highlights the need for students to find job opportunities, with standard options being domestic and hotel jobs. Another significant reason for leaving is education, as some students seek better quality education or higher levels of education, which traffickers exploit.

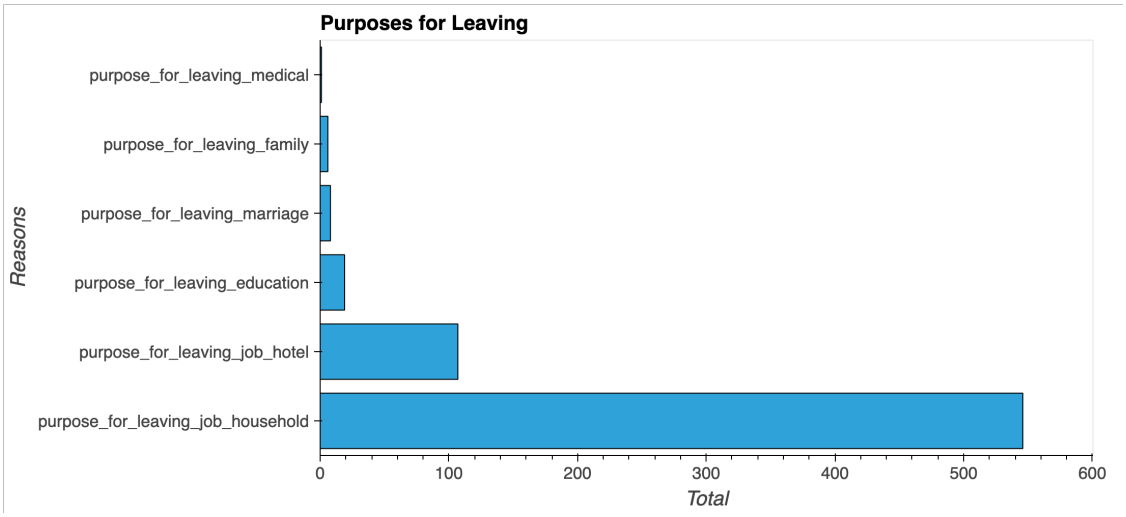


**Figure 4.18:** Uganda Top Reasons for Leaving

The data shown in Fig. 4.19 reveal that the most common forms of exploitation experienced by victims are forced labour, physical abuse, and sexual abuse. Primary and secondary school students in Uganda are highly vulnerable to trafficking and various forms of exploitation. Victims are primarily exploited within Uganda, and fewer are detected in Kenya and Malawi.

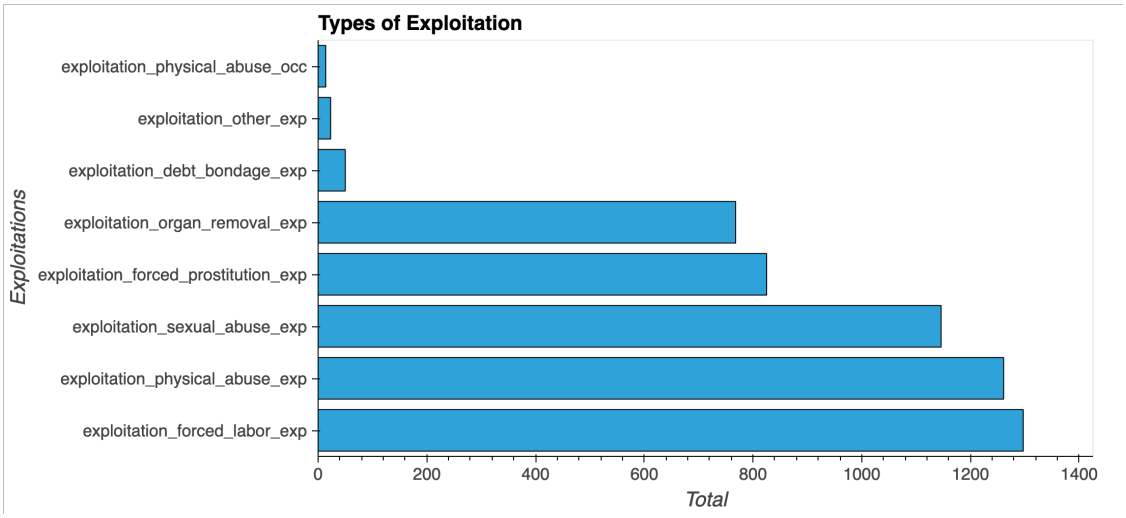


**Figure 4.19:** Uganda Top Forms of Exploitation

### Malawi

Malawi is another country with a high number of trafficking victims. Similarly to Uganda, most of the victims are secondary students. However, unlike Uganda, there are more male victims than females. Primary students are also highly targeted by traffickers, unlike university students, as shown in Figure 4.20. Malawi's education system faces challenges such as limited infrastructure and a shortage of teachers, especially in primary schools, leading to poor quality education and high dropout rates (Starfish Malawi, n.d.; Cross Catholic Outreach, 2021). These circumstances increase the risk for children to be subjected to various forms of exploitation, including forced labour, physical violence, and sexual exploitation (Cross Catholic Outreach, 2021).

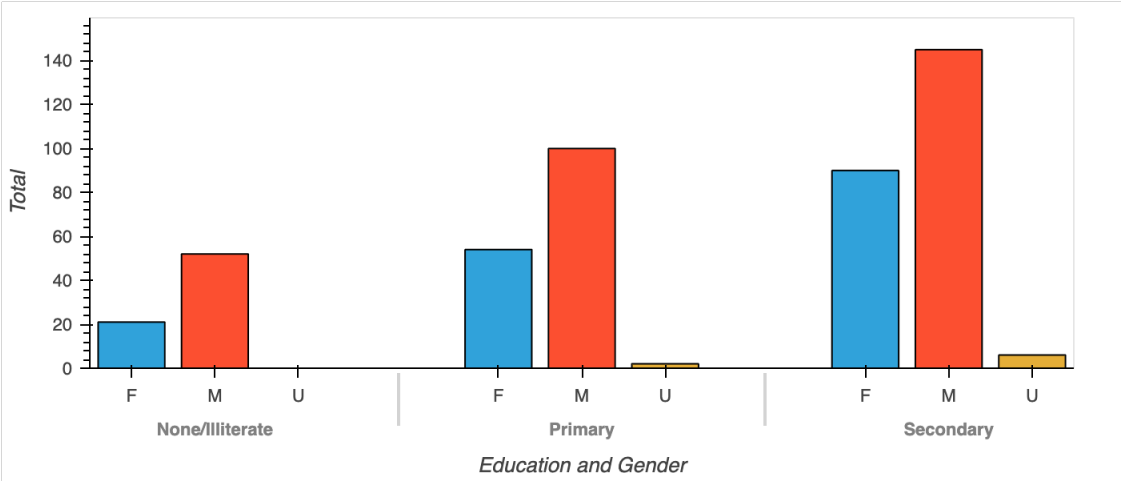


**Figure 4.20:** Malawi Education Statistics

According to Figure 4.21, most of the victims in Malawi are farmers, followed by students. Agriculture is the most prevalent industry in Malawi, with more than 85% of the population dependent on it and a labour participation rate of 55%, which explains the high proportion of farmers among the victims (Botha, n.d.). Similarly to Uganda, students are vulnerable to trafficking, raising concerns that traffickers can potentially recruit in or near schools.

Figure 4.22 illustrates that victims in Malawi are predominantly recruited through personal networks, including family and friends, followed by direct physical

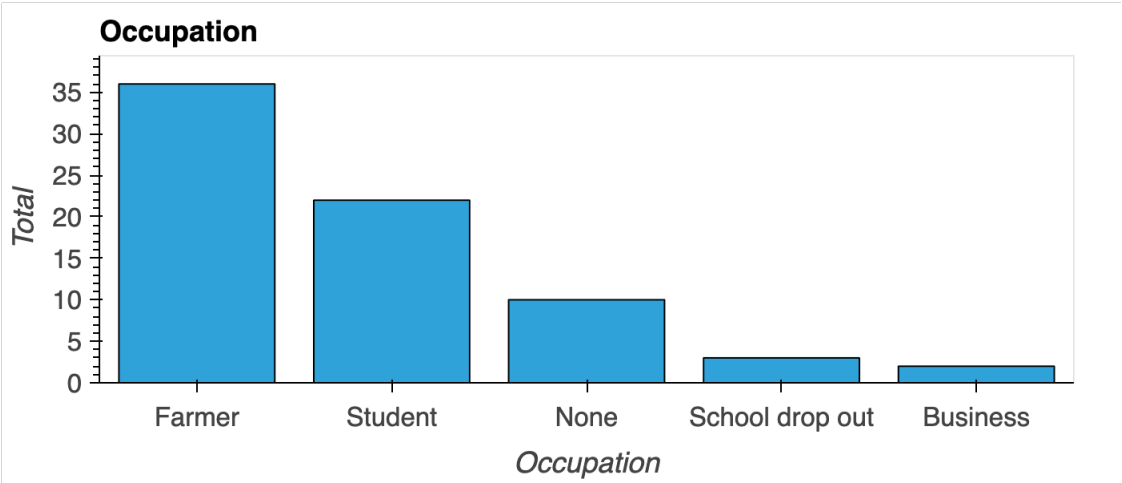


**Figure 4.21:** Malawi Top Occupations

and online approaches by the suspected trafficker, as shown in Figure 4.22. Online recruitment was primarily done through social media.

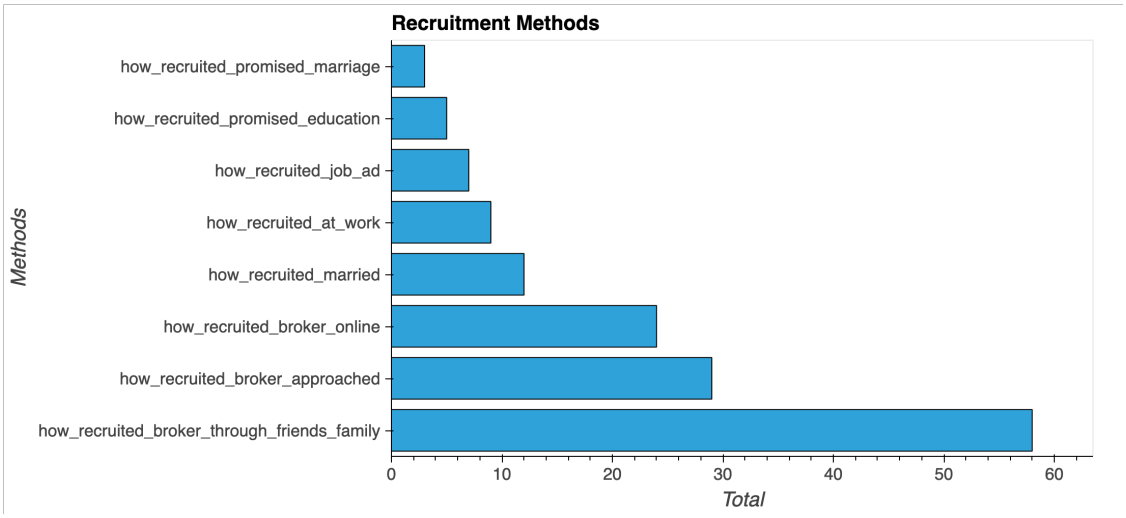


**Figure 4.22:** Malawi Main Recruitment Methods

Like Uganda, Malawians are targeted for employment opportunities in households and hotels, as seen in Figure 4.23. Another common reason for recruitment is marriage, which is more prevalent among female victims. Reports indicate that Malawian girls often drop out of school for early or forced marriages, with 42% getting married before age 18, according to the Johns Hopkins Centre for Communication Programs (Desmon, 2022). The average age of the victims in Malawi is 20 years.

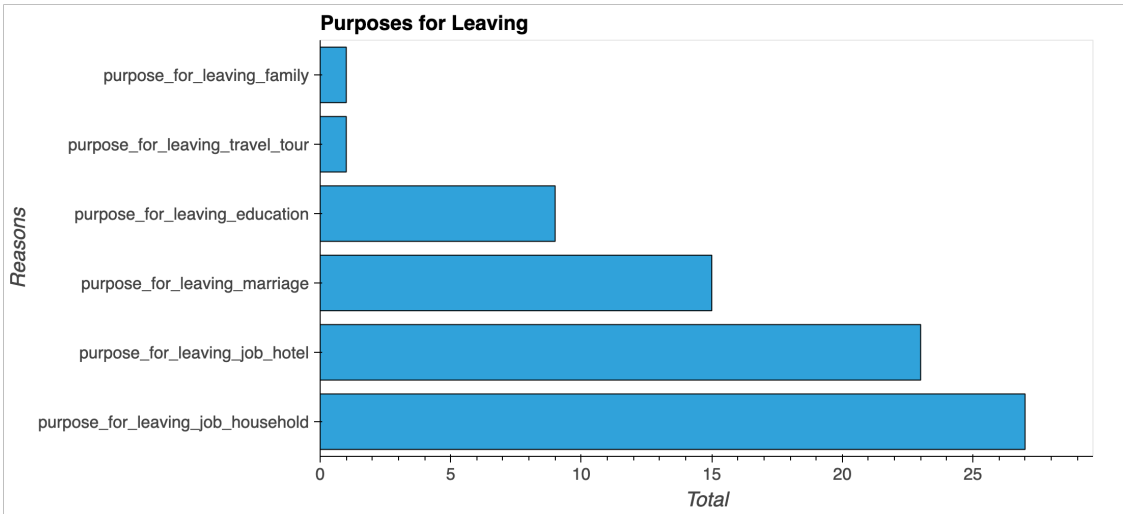


**Figure 4.23:** Malawi Top Reasons for Leaving

There is a typical pattern between Malawi and Uganda, where victims are targeted for forced labour, physical violence, sexual abuse, and forced prostitution, as seen in Figure 4.24. Most Malawian victims were recruited within the country and only a few were recruited to South Africa. Malawi shows a slight increase in the number of victims, indicating that it also serves as a destination country. However, Malawi is not economically stable, so other factors may contribute to making it a suitable destination or transit country for human trafficking.

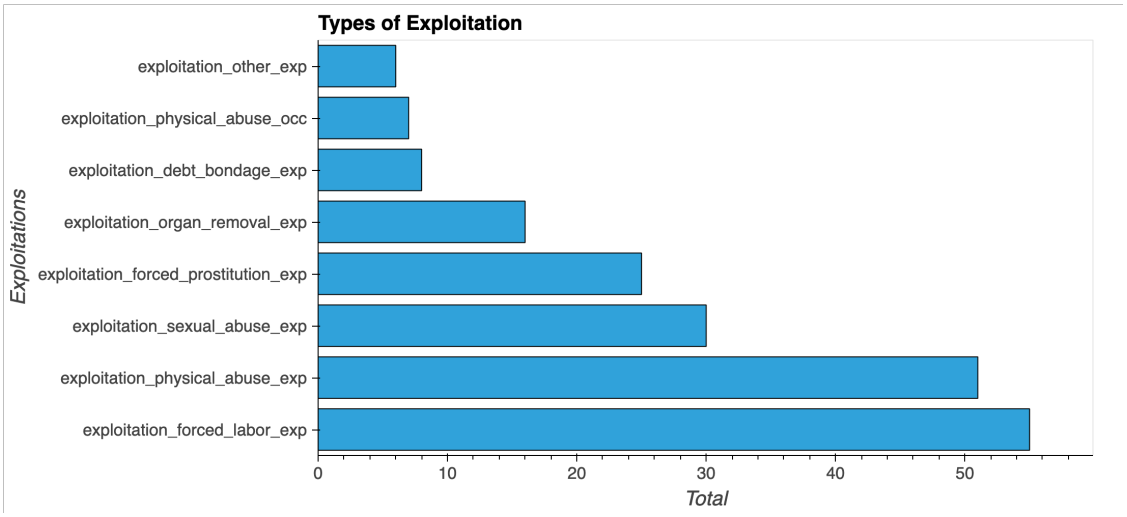


**Figure 4.24:** Malawi Top Forms of Exploitation

### South Africa

The dataset indicates that South Africa has the third highest number of victims. As shown in Figure 4.25, most of the victims in South Africa are women and predominantly have a secondary or university education.

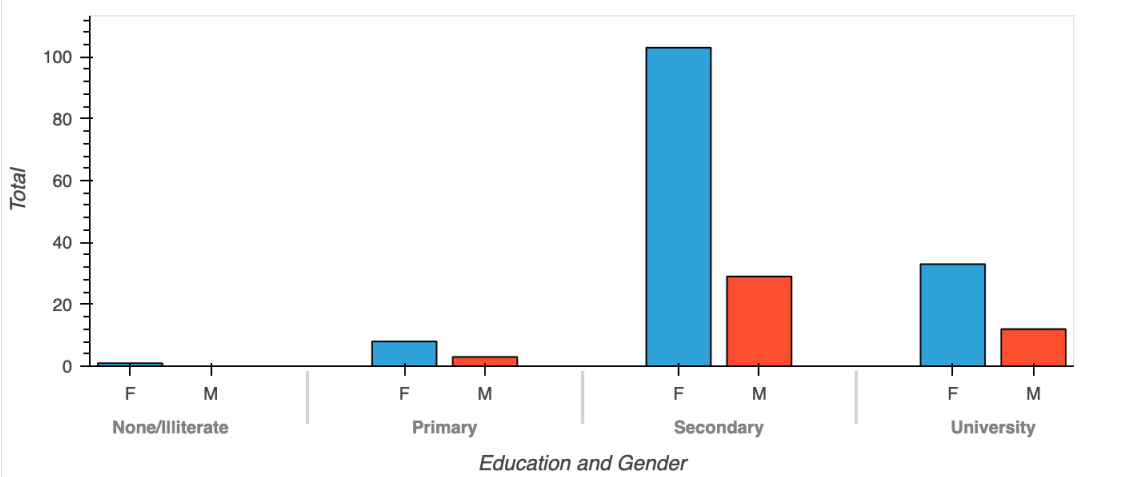


**Figure 4.25:** South Africa Education Statistics

Furthermore, Figure 4.26 shows that a significant proportion of the victims in South Africa are unemployed. Despite being one of the most developed countries on the continent, South Africa is grappling with a high level of unemployment. South Africa currently faces an unemployment rate of 32.6%, the lowest since 2021 (Gumbi, 2023; Trading Economics, 2023a, 2023b). The youth unemployment rate is even more alarming, at 60.7% (Trading Economics, 2023a). The average age of the victims in South Africa is 28, which is a much older population compared to Malawi (20) and Uganda (17).

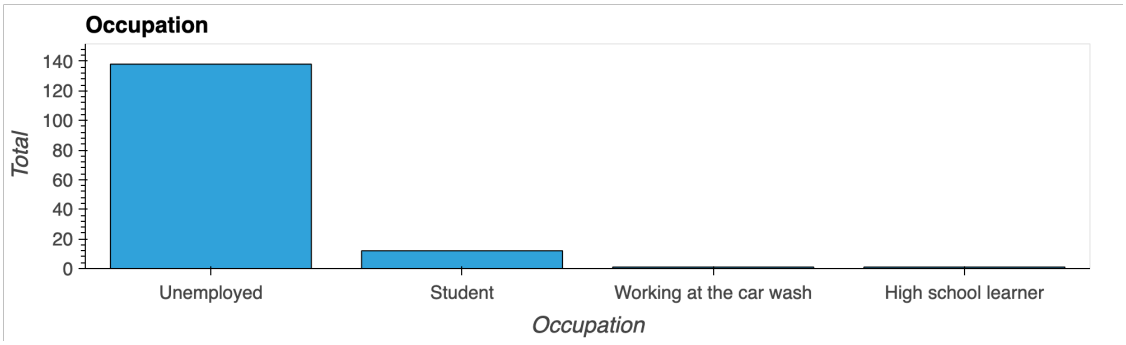


**Figure 4.26:** South Africa Occupations

On the other hand, Figure 4.27 reveals that the most common recruitment method in South Africa is through **job advertisements**, followed by **approaches made through brokers**, online and offline.

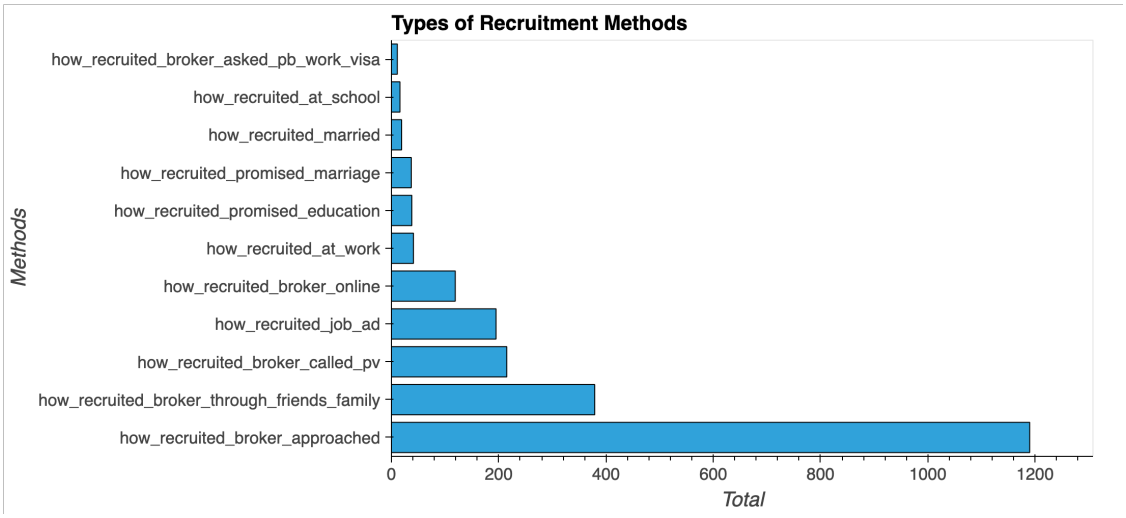


**Figure 4.27:** South Africa Main Recruitment Methods

Similarly to Uganda and Malawi, most victims in South Africa were recruited for household jobs, followed by hotel jobs, as shown in Figure 4.28. A significant trend is that most victims did not have identification during interception. People from less developed countries such as Zimbabwe, Malawi and Lesotho often travel to South Africa, both legally and illegally, due to its higher level of development (Onuoha, 2011). This could contribute to the lack of identification among victims, as migrants are often at higher risk of human trafficking.

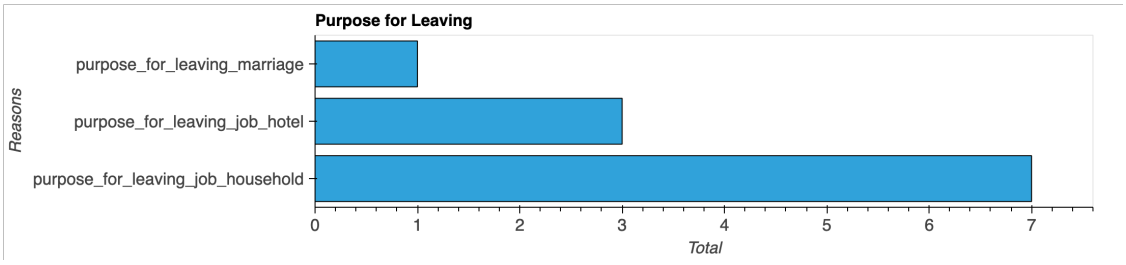


**Figure 4.28:** South Africa Top Reasons for Leaving

Furthermore, most victims are exploited for physical abuse, forced labour, and sexual abuse, as seen in Figure 4.29.

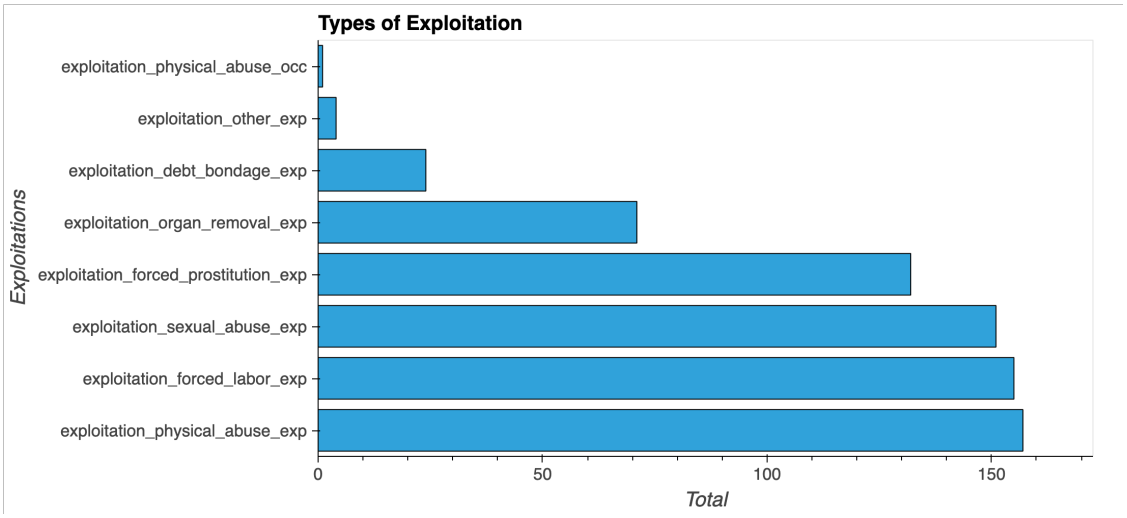


**Figure 4.29:** South Africa Top Exploitations

### Kenya

Similarly to the countries mentioned above, most victims are secondary students. The image 4.30 summarises the country's education statistics, revealing that secondary education is prevalent, closely followed by primary school enrolment. Primary education in Kenya is free and compulsory, a policy that has significantly increased student enrolment in primary schools (Embassy of the replublic of Kenya, n.d.; Britannica, 2024). In addition, the government has implemented programmes to facilitate students' transition to secondary schools by offering subsidised fees (Embassy of the replublic of Kenya, n.d.).

However, as UNICEF pointed out, the challenges within the Kenyan education system affect the quality of education (UNICEF, 2023). Issues such as a shortage of quality teaching and overcrowded classrooms affect students' learning experiences. Furthermore, specific communities experience high dropout rates due to a lack of interest in school, long distances to educational institutions, and early marriages (UNICEF, 20230. These challenges within the education system can potentially create vulnerabilities that traffickers might exploit when recruiting students, both in secondary and primary schools (UNICEF, 2023).

The image represented by 4.31 provides a breakdown of the occupations of the victims, and, notably, the largest group consists of students, followed by domestic

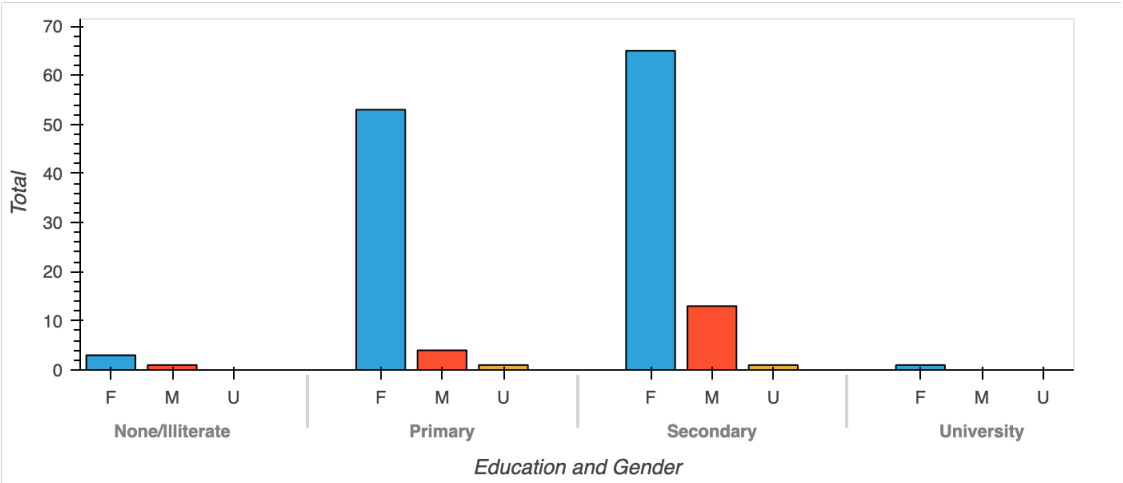


**Figure 4.30:** Kenya Education Statistics

workers. Kenya has a significant number of domestic workers, estimates stating that around 2 million people are employed in this sector (Nzamba Kitonga Advocates LLP, n.d.). The average age of the victims is approximately 19 years old.

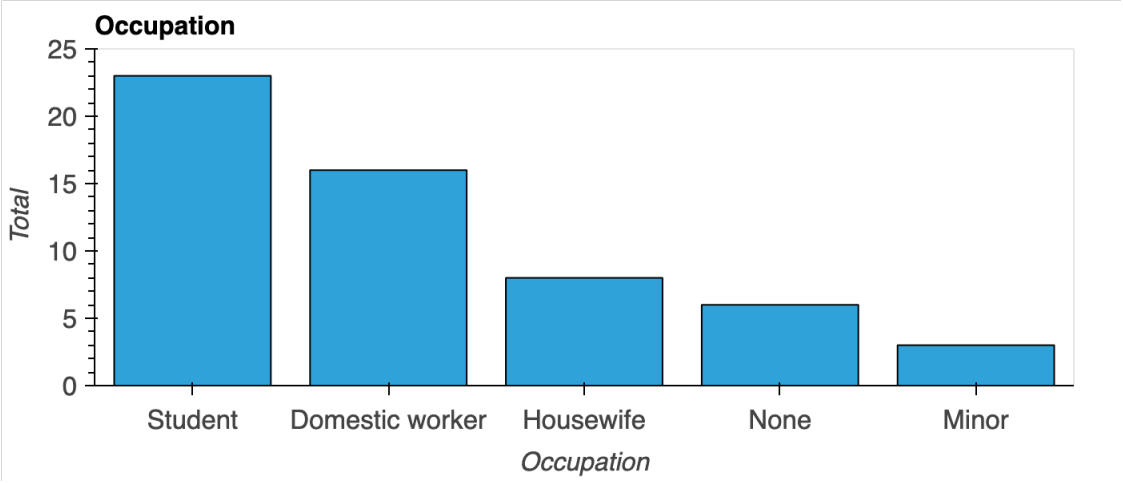


**Figure 4.31:** Kenya Occupations

As depicted in Image 4.32, it is evident that the primary method of recruitment for victims in the country is through family and friends, followed by online recruitment.

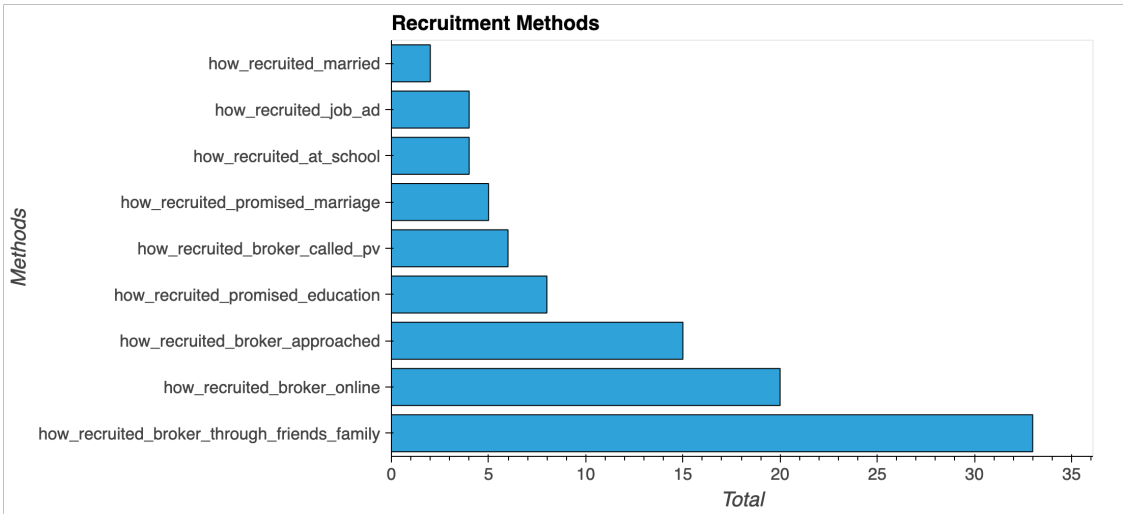


**Figure 4.32:** Kenya Recruitment Methods

Similarly to the countries above, the primary reasons for recruitment in Kenya are jobs in the home and hotel and educational opportunities, as shown in Figure 4.33. Kenya places a significant value on education (Britannica, 2024) and this emphasis on education likely motivates people to seek opportunities to advance their education.

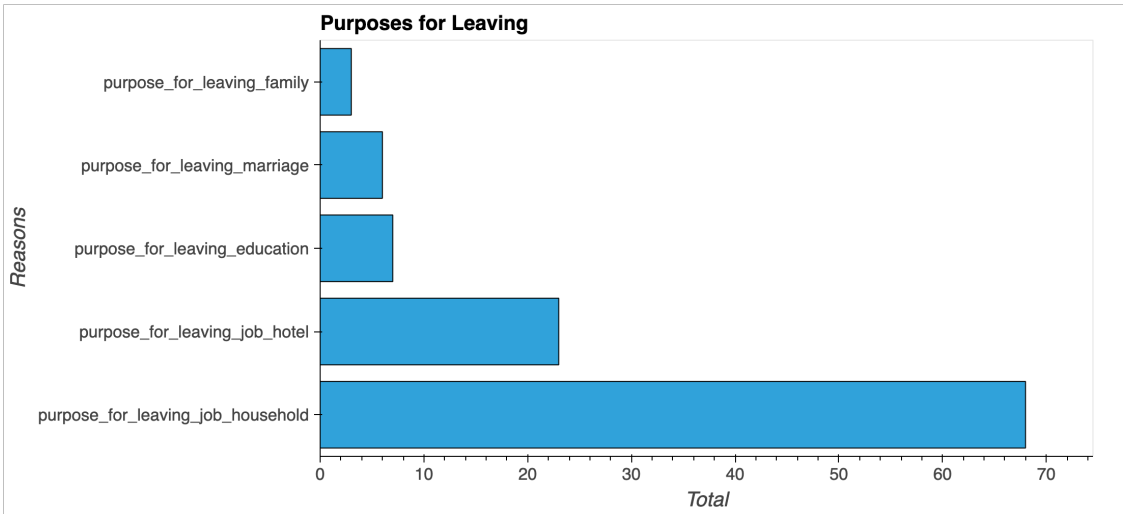


**Figure 4.33:** Kenya Purpose for Leaving

Kenyan victims are mainly exploited for forced labour and sexual abuse.

### Angola

There was a sharp increase in the number of trafficking victims in Angola in 2022. Most victims are male and illiterate, as seen in Figure 4.34. The relatively low

literacy rate in Angola (Macro Trends, 2023) also indicates that better education programmes may be needed to prevent people from falling victim to trafficking.

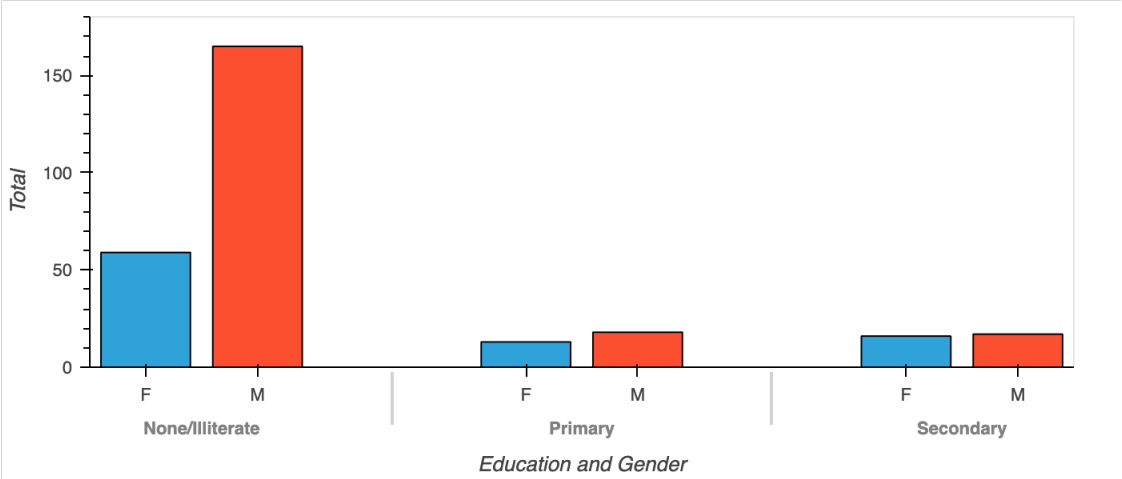


**Figure 4.34:** Angola Education Statistics

Individuals without any education in Angola are more susceptible to trafficking, unlike Uganda, South Africa, and Malawi, as shown in Figure 4.35. More work is needed to investigate why students are more prone to trafficking in the other countries, unlike in Angola.




**Figure 4.35:** Angola Top Occupations

In Angola, victims were predominantly recruited by friends and family, and many actively sought job opportunities that put them in contact with traffickers. This trend is likely because many victims were unemployed and vulnerable to

exploitation. Forced labour, physical abuse, and sexual abuse were reported as the primary forms of exploitation.

The data revealed that household jobs are the main reason for leaving among Angolan victims, followed by travel tours, as illustrated in Figure 4.36. Furthermore, most Angolan victims are intercepted in Namibia, suggesting that traffickers frequently promote Namibia as a destination country. All but one victim were identified in Namibia in this dataset.

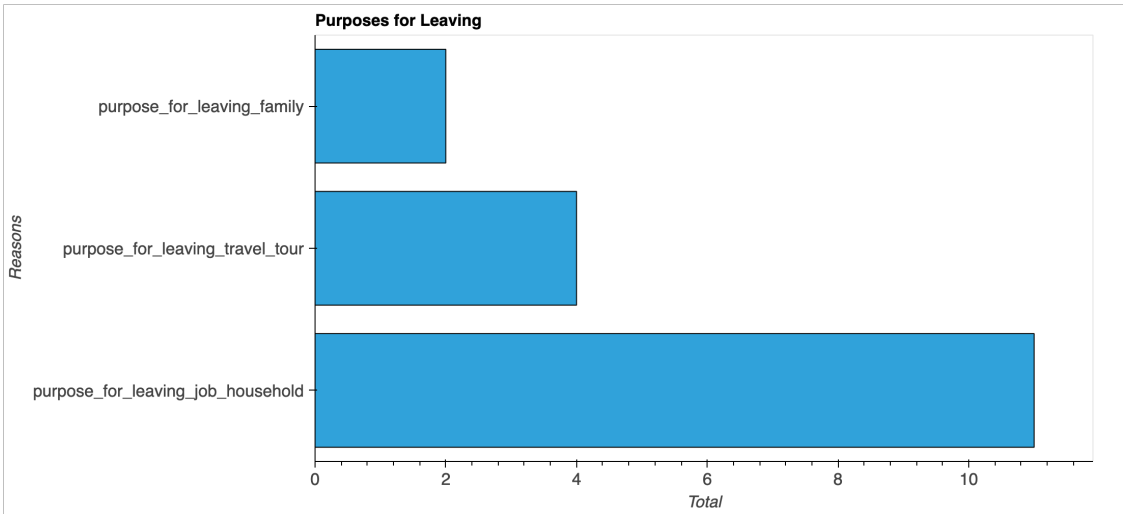


**Figure 4.36:** Angola Top Reasons for Leaving

### 4.6.3 General Hypothesis Testing

In this section, a statistical analysis is also performed on the dataset to draw conclusions about recruitment and further validate the existing assumptions obtained from the analysis and previous research.

Two hypotheses are tested using the null hypothesis ($H_0$) versus alternative tests ($H_1$). Similarly to the previous section 4.5.2, CHI test and residual analysis are used to analyse the data(Vaidya, 2023). Similar suitability tests are also performed before proceeding with the CHI test.

**Is there a relationship between gender and recruitment through promised jobs?**

This hypothesis aims to evaluate whether individuals of a specific gender are more susceptible to recruitment through fraudulent job opportunities. While recruitment through fake relationships has predominantly been associated with females Anthony, 2018, there is limited literature exploring gender differences in recruitment through deceptive employment schemes. This study seeks to identify which gender is more vulnerable to such practices within the African context. Therefore, the relationship between gender and recruitment by promise of a job is examined.

$H_0$: There is no relationship between gender and recruitment through the promise of a job

$H_1$: There is a relationship between gender and recruitment through the promise of a job

Based on the test results, which included a chi-square value of 43.4, an extremely small p-value of 3.84e-10, and a degree of freedom of two, therefore, the null hypothesis is rejected. There is a significant relationship between gender and recruitment through the promise of a job.

Further analysis was performed to compare the different categories of variables, revealing that women are more likely to be recruited through job promises. This analysis used residual analysis to determine statistical significance. The gender category has the following fields: females, males, and undefined.

**Does education play a role in whether individuals are recruited through an agency?**

The following hypothesis explores the relationship between education level and recruitment through employment agencies. Preliminary analysis suggests that traffickers often exploit agency channels to recruit individuals seeking job opportunities. This hypothesis seeks to determine whether individuals with specific levels of education are more likely to be targeted or recruited through mployment recruitment agencies.

$H_0$: There is no relationship between education and recruitment using an agency

$H_1$: There is a relationship between education and recruitment using an agency

The test results produced a chi-square value of 114.9, a p-value of 9.8434e-25, and three degrees of freedom. These findings indicate a significant relationship between education and recruitment through agencies, leading to the rejection of the null hypothesis.

Further analyses comparing different education categories within the variables revealed that this relationship was primarily observed among university students. These analyses used residual analysis. The test results suggest that university students or graduates are more susceptible to recruitment through agencies than other groups, followed by primary school students.

## 4.7 Discussion

This section discusses the findings of the previous sections, explores the design implications based on these findings, and outlines the plans for the next tasks in this thesis.

### 4.7.1 Dataset 1

Dataset 1 provides data on individuals intercepted in various African countries during human trafficking recruitment. During the past five years, a positive correlation was observed, particularly in 2022. One potential event that could have contributed to increased victim trafficking in 2022 was the Ukrainian war. The World Bank reported that this conflict had significant consequences for global economies in Africa, leading to high inflation rates and food insecurity, which negatively affected approximately 140 million people in 2022 (The World Bank, 2022). Economic hardships such as poverty and food insecurity often drive people to seek better opportunities for a more comfortable life, making them vulnerable to exploitation by traffickers. However, the United Nations has also reported that the majority of conflicts originate in Africa and the Middle East, which consequently results in the highest proportion of victims originating in and migrating to these regions (United Nations: Office on Drugs and Crime, 2023).

The dataset revealed that the majority of the victims were **children and youth**. This finding is consistent with reports that child trafficking is a major problem in Africa (United Nations Office on Drugs and Crime, 2023). A report states that more than 200,000 children in West Africa are recruited annually and transported from these countries for various forms of exploitation, including prostitution and domestic slavery (Allais, 2006). It is critical to implement strategic measures aimed at protecting children and youth to prevent them from being victimised by trafficking.

These results emphasise the need to **increase awareness** of human trafficking **in transportation hubs** such as airports, bus and train stations. It is recommended that stakeholders employ resources such as posters to disseminate information about human trafficking, its dangers, and ways to access support. Tangible resources should be strategically placed in areas that victims are likely to encounter (Simonson, 2021). Viewing this information can encourage victims to seek help from anti-trafficking stakeholders at these transportation hubs, such as law enforcement and non-governmental organisations such as Organisation A.

Providing staff, such as airport and airline personnel in transport hubs, with **training to identify indicators** of human trafficking can significantly improve efforts to detect and prevent this form of exploitation. Previous research has also indicated that essential anti-trafficking stakeholders and front-line workers lack adequate training to identify human trafficking (Ediae and Kuteesa, 2024). Enhanced training can lead to an increase in identified cases, as many suspected instances may be overlooked by Organisation A due to limited personnel and the expansive nature of airports and stations. The training should focus on identifying vulnerabilities that Organisation A seeks when interviewing suspected victims. These vulnerabilities can be summarised into six main themes, which are described below. The terms "victim" and "trafficker" here refer to suspected victims and traffickers.

1. **Insufficient knowledge of travel arrangements**

    - The victim does not have information on their destination.
    - This is the first international trip of a victim.

- Family and friends arranged travel for the victim.
- The victim only has a one-way ticket.

2. **Language and communication barriers**

- The victim does not speak the language of the destination country.
- The victim does not speak English.

3. **Suspicious travel arrangements**

- The victim does not have a visa to the destination country.
- The flight was paid in cash.
- A non-family member paid for the trip.
- The victim has insufficient resources for travel.

4. **Vulnerable relationships and dependency**

- The family does not know about the victim's travel plans.
- Minor travelling without a guardian.
- There is no clear connection between the victim and the suspected trafficker.

5. **Online and unfamiliar connections**

- The victim found a job opportunity on social media.

6. **Suspicious Behaviour**

- The victim is visibly afraid of the suspected trafficker.
- The victim refuses to provide family information.
- A group of potential victims are travelling together, but have never met before.
- Someone speaking on behalf of a victim.

### 4.7.2 Dataset 2

Dataset 2 provided information on several aspects of the trafficking process, such as recruitment techniques, occupations, factors that lead to recruitment and different forms of exploitation. Furthermore, data from multiple countries were examined to gain a more comprehensive perspective at a granular level.

According to the data, most of the trafficking victims have completed secondary education and in some countries such as Uganda they are currently enrolled in school. This finding raises questions about why secondary school students and graduates are particularly susceptible to trafficking. More research is needed to determine the underlying reasons for this. It is also evident from the analysis that traffickers frequently target individuals, such as students directly. This may be related to factors such as the teacher-to-student ratio in secondary schools in Africa (UNESCO, n.d.), which can result in fewer students being supervised. Since low-income students can now attend secondary schools due to their free education in countries such as Kenya and Malawi (UNICEF, 2023), traffickers can take advantage of this opportunity by approaching them in schools. This could explain why many of the victims in the dataset were current primary or secondary students.

In addition, some countries had a high number of university students, particularly women, who were recruited, and there was a noticeable trend of debt bondage among them. It is essential to discourage students from obtaining loans from individuals and instead encourage them to acquire loans through official loan bodies. Similarly, targeted prevention and awareness campaigns are needed in higher education institutions. These campaigns can be integrated into orientation sessions or tailored for students who face financial difficulties. University students are also susceptible to fraudulent recruitment agencies, so stakeholders should compile and raise awareness of verified and safe recruitment agencies for university students and others seeking jobs through these agencies.

Educators should incorporate awareness programmes as a component of school curriculums across all levels of education, with the aim of allowing students to recognise fraudulent opportunities and understand the consequences associated with

them. The US Department of State has emphasised the need to empower vulnerable individuals with the ability to recognise fraudulent opportunities (Department of State, 2023). In addition, improving school education standards can also be beneficial in tackling human trafficking.

Stakeholders should also involve local communities to raise awareness of crime and the tactics involved. By collaborating with community leaders and organisations, people, especially in low-income areas, can be informed and vigilant.

Governments play a crucial role in addressing the factors that contribute to increased trafficking on the continent. It is essential to address the root causes of trafficking, such as unemployment and poverty. Unemployment is a significant factor that presents numerous opportunities for trafficking. The research findings align with reports from the United Nations, which state that victims are frequently lured into domestic work, cleaning, and bar work (United Nations: Office on Drugs and Crime, 2023). Failure to tackle these core issues leaves individuals vulnerable to traffickers.

The findings also indicate that leaving household jobs is prevalent, necessitating interventions to address this problem. There is a need to establish safer avenues for people to seek employment opportunities, especially domestic and hotel jobs. Ensuring the safety of job platforms, both online and offline, requires a concerted effort from various stakeholders. Governments play a significant role in regulating online job posts. It is essential to hold online platforms responsible for the information shared on their platforms so that they also make efforts to combat online recruitment. Additional regulation efforts can involve flagging or banning recruitment agencies and online account profiles that facilitate trafficking.

### 4.7.3 Design Implications

These findings have produced several design implications for technology solutions that help combat human trafficking.

**Targeted Awareness Campaigns -**

Designers of awareness campaigns aimed at combating human trafficking in Africa should leverage online platforms, particularly those popular with children and young people, to deter recruitment efforts. Tailoring campaigns to effectively engage these vulnerable populations is crucial, as conventional methods may not be suitable. Innovative strategies, such as targeted social media campaigns or advertisements that have proven effective in reaching younger audiences, should be implemented (Horowitz Research, 2023).

Targeted campaigns increase the likelihood that the audience will engage with the content. Furthermore, incorporating reward-based mechanisms can motivate users to interact with and share campaign materials on social media platforms (Shawky et al., 2019). This approach not only broadens the reach of the campaign, but also fosters community-driven efforts to combat trafficking.

For young audiences, incorporating interactive and relatable elements, especially those created by their peers, as seen in successful public health initiatives, could further enhance engagement and message retention (World Health Organisation, 2023).

Finally, campaign content should be designed to be easily understood by people of different backgrounds, helping potential victims identify suspicious activities, verify opportunities, and access support from anti-trafficking organisations.

**Border Security Systems -**

Currently, the US and the European Union use AI management systems to monitor their respective borders (Tyler, 2022). Implementing similar AI-based systems in various transportation hubs in Africa could help detect potential trafficking victims. Countries in Sub-Saharan Africa are known to detect far fewer victims compared to other regions, raising concerns (United Nations: Office on Drugs and Crime, 2023). AI systems like these would be beneficial to Organisation A and other organisations in improving victim detection.

When designing image and video recognition systems for border security purposes, it is important to design algorithms that operate in real time and assist anti-trafficking stakeholders, such as Organisation A, in conducting investigations at transportation hubs and intercepting victims and traffickers. These systems should possess the ability to analyse visual data to identify passengers who meet specific criteria, such as unaccompanied minors or individuals travelling without luggage to known trafficking hotspots.

The design of these technologies must also consider their accuracy, privacy, and ethical implications. The European Union's trials of systems such as iBorderCtrl highlight the need to carefully consider these factors to prevent false positives, discrimination, and potential violations of individual rights. Several countries in the European Union conducted iBorderCtrl trials, which used image recognition technology to evaluate microexpressions with the aim of detecting deception during interviews (Tyler, 2022). However, due to concerns about its accuracy, privacy and ethical implications, the system was not implemented beyond the trial periods.

**Developing Discrete Alert Systems -**

The findings indicate that victims sometimes approach Organisation A and other officials for help. Moreover, the United Nations reported that most victims initiated their own efforts to seek assistance from trafficking situations, highlighting the need for systems that allow victims to seek help (United Nations: Office on Drugs and Crime, 2023). Stakeholders should consider using online and physical methods, such as Wi-Fi network landing pages and physical systems in restrooms. However, it is essential for stakeholders to provide discrete methods for victims to seek help without alerting their traffickers, as it can pose a threat to their lives (Simonson, 2021).

Wi-Fi network-based alert systems can enhance awareness of anti-trafficking initiatives in transportation hubs and provide a discreet means for victims to request help. The design of landing pages should offer clear and accessible information on seeking assistance, along with a subtle "Help" option that allows potential victims to alert authorities without drawing attention.

For physical systems in restrooms, stakeholders can use AI technologies to allow victims to seek help and share their data through manual or biometric means. The design must prioritise user-friendliness to ensure that victims can easily navigate the system under pressure, with minimal steps required for swift access to assistance.

**Developing Anti-trafficking Measures to Address Fake Opportunities -**

The findings confirm that fake opportunities are a common method of recruiting victims in Africa. A software solution can be designed to detect suspicious opportunities, such as job offers or scholarships, by using indicators identified in this chapter, such as the promise of high-paying jobs. The US Department of State emphasises the need for governments to develop safe avenues to verify job opportunities (Department of State, 2023).

A software tool can be developed to help Organisation A personnel investigate opportunities during interception. This tool can verify the legitimacy of opportunities, reducing the time that staff spend enquiring about the validity of a job or scholarship. For example, the software can consider various factors identified as important in the analysis, such as salary, victim qualifications, and the authenticity of the job posting. However, more research is necessary to identify additional indicators on scholarship and job opportunities before developing a software solution. The following section discusses how the next chapter focusses on identifying additional indicators of fraudulent jobs.

### 4.7.4 Direction of Thesis

There is a notable knowledge gap on the tactics used by traffickers to recruit victims on digital platforms in Africa. The dataset analysis did not provide adequate information on these indicators and an overall understanding of fraudulent jobs on digital platforms. The findings in this chapter influence the direction of the thesis in several key ways:

1. The findings have demonstrated a tendency for traffickers to advertise non-existent employment opportunities on social media platforms and that cases

involving such purported positions often possess sufficient evidence to be categorised as human trafficking. The subsequent phase of this research will involve a collaborative effort with Organisation A to investigate fraudulent job advertisements for human trafficking on online platforms. The primary objective will be to understand the strategies employed by traffickers on social media and to identify the key indicators of these deceptive practices.

2. This thesis initially focused solely on sex trafficking; however, the tactics employed to recruit individuals are similar in most forms of human trafficking. Consequently, the findings are applicable not only to sex trafficking but also to other types of human trafficking. Furthermore, Organisation A indicated that they are unable to differentiate between recruitment strategies used for the different types of trafficking during investigations and consequently do not focus on a specific category of human trafficking.

3. The knowledge acquired in this chapter and in the subsequent chapter will facilitate researchers in designing a study that uses AI to identify fraudulent job advertisements. The indicators of fraudulent employment opportunities identified in this chapter and in Chapter 5 inform the queries that will be employed to collect data in Chapter 6.

## 4.8 Conclusion

This chapter has provided an analysis of two datasets and a review of the relevant literature, providing valuable information on the multifaceted issue of human trafficking recruitment in various African countries.

The first dataset allowed us to gain insight into the interception process and the criteria used by Organisation A to identify high-risk trafficking situations. This included an examination of victim profiles, industries associated with trafficking, and signs of vulnerability. In addition, the dataset sheds light on the tactics used by traffickers, such as marriage and employment promises, and the vulnerabilities that highlight trafficking.

The second dataset enhanced the understanding of recruitment methods, victims' educational levels, occupations, reasons for recruitment, and the common types of exploitation. It offered a detailed analysis of five countries with a high number of victims, showcasing the specific tactics used in these regions.

Hypothesis testing was also performed on both datasets to assess relationships within the data. These tests helped us identify significant associations and further investigate the relationships between different categories within the data.

An overarching theme of these datasets is the need to address systematic issues in African countries, such as unemployment, poverty, and conflicts, as these factors contribute significantly to human trafficking. Traffickers exploit these vulnerabilities, particularly among children and youth.

# 5

# Investigating Fraudulent Job Advertisements for Human Trafficking Online

## Contents


**5.1 Introduction . . . . . 160**
**5.2 Methodology . . . . . 161**
5.2.1 Participant and Platform Background . . . . . 162
5.2.2 Online Observation: Social Media Groups . . . . . 164
5.2.3 In-person Observation . . . . . 165
5.2.4 Semi-structured Interviews . . . . . 166
5.2.5 Case Notes . . . . . 167
5.2.6 Analysis . . . . . 167
5.2.7 Ethical Considerations of Digital Observations . . . . . 168
**5.3 Findings . . . . . 169**
5.3.1 Methods Traffickers Employ to Recruit Victims on Social Media . . . . . 169
5.3.2 Organisation Work: Raising Awareness . . . . . 170
5.3.3 Organisation Work: Tackling Online Human Trafficking Recruitment . . . . . 172
5.3.4 Identifying Suspicious Job Advertisements . . . . . 177
5.3.5 Characteristics of Suspicious profiles . . . . . 184
**5.4 Discussion . . . . . 184**
5.4.1 Summary . . . . . 184
5.4.2 Analysis of Organisational Work . . . . . 185
5.4.3 Opportunities for the Use of Technology to Identify and Investigate Suspicious Job Posts . . . . . 188
5.4.4 Design Implications for Social Media Platforms . . . . . 192
**5.5 Reflection on Fieldwork . . . . . 194**

**5.6 Conclusion . . . . . 195**

## 5.1 Introduction

This chapter addresses the third research question: *What are the indicators of fraudulent job advertisements associated with human trafficking recruitment on online platforms, and how do anti-trafficking stakeholders effectively identify these indicators?* By addressing this research question, the study sheds light on (i) the methods used by traffickers to recruit victims on social media through fraudulent job offers, (ii) the characteristics of job postings used for online human trafficking recruitment, and (iii) the strategies employed by anti-trafficking organisations to investigate job advertisements related to human trafficking recruitment on online platforms.

This chapter builds on the ongoing collaboration with Organisation A discussed in Chapter 4. However, the focus is on one African country because, as previously mentioned, among the countries where the organisation operates, this is the only country that investigates fraudulent employment opportunities on social media platforms. As in Chapter 4, the organisation will remain anonymous to protect its identity. The team in this country monitors and investigates activities on a single platform, which is the focus of this study. This platform is also widely used in other countries, both in the Global North and in the Global South, and is known to facilitate human trafficking recruitment based on previous literature (Dorton, 2019; Latonero and Movius, 2011; Progetto Tenda et al., 2021; Anthony, 2018; Feehs and Wheeler, 2020; United Nations Office on Drugs and Crime, 2020)and the analysis conducted in Chapter 4. The platform is also anonymised because of concerns that traffickers may alter their tactics if this research is discovered. This social media platform of focus will be referred to as Platform A, the country will be designated as Country A, and the organisation operating in Country A will be termed Organisation B.

To address the research question and achieve the objectives of this chapter, multiple methods were employed: on-site observations of Organisation B's operations were conducted in Country A; previous cases related to human trafficking were analysed; semi-structured interviews were conducted with experts in Organisation B; and observations of job advertisement groups on Platform A were performed. Through these research methods, this study makes the following contributions:

1. Insights into the tactics employed by 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 organisations to investigate human trafficking recruitment on social media platforms, as well as the challenges they face in their current setup.

3. Insights into the opportunities of using technology to combat human trafficking recruitment.

4. Design implications for features that could be implemented on social media platforms to help reduce the likelihood of human trafficking recruitment.

The remainder of this paper is organised as follows. Section 5.2 discusses the methods employed to achieve research objectives. Section 5.3 presents the findings and Section 5.4 discusses these findings and their design implications. Section 7.4.3 addresses the limitations and challenges of this research, and Section **??** concludes the chapter and outlines the next steps of this thesis.

## 5.2 Methodology

Figure 5.1 shows an overview of the research process. The initial phase commenced with the observation of job advertisement groups on Platform A in collaboration with Organisation B, followed by in-person observations conducted in Country A with Organisation B, concurrently with the semi-structured interviews. The following phase involved an analysis of existing cases that had been investigated

on Platform A and found to be associated with human trafficking. The final phase encompassed the analysis of all data collected from the preceding methodologies.

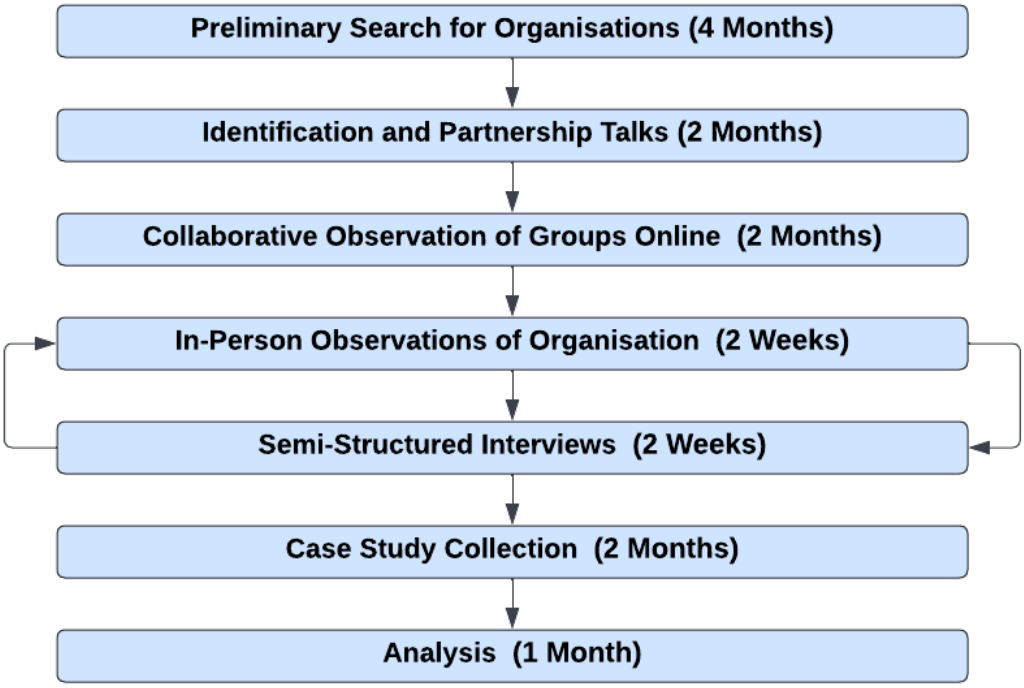


**Figure 5.1:** Chapter 5 Methodology Process

The following section provides background information on Organisation B and the participants involved in the study.

### 5.2.1 Participant and Platform Background

#### Organisation B and Research Participants

Organisation B implements a multifaceted strategy aimed at raising awareness among potential victims about the risks associated with human trafficking and investigating fraudulent opportunities for human trafficking on Platform A. The primary objective is to prevent potential victims from pursuing high-risk opportunities that can result in them being trafficked. In contrast to other countries where Organisation A operates, Organisation B seeks to intercept victims prior to their transportation. It initiates various investigative processes to determine whether job advertisements

are related to human trafficking, and subsequently contact potential victims to prevent them from proceeding with the job.

Organisation B operates in a single city and maintains a physical office where its operations are headquartered. These offices are located in a church compound, generously provided to them free of charge. However, during the observation period, the project manager and investigators frequently worked from various locations within the city to carry out awareness campaigns about human trafficking. Due to time and cost constraints, they carried out investigations on Platform A in the same locations where awareness campaigns were conducted. These locations were typically low-income neighbourhoods where people who are more vulnerable to human trafficking can usually be found.

Organisation B comprises a diverse group of individuals, with the main points of interaction in this study being the project manager, the director, the data analysts, and the investigators. The investigators and project manager are at the forefront and are responsible for investigating cases of human trafficking on Platform A. The researcher engaged with five investigators who had eight months and two years of experience working on human trafficking within the organisation, a project manager with seven years of experience, and a director with five years of experience.

### Platform A

Platform A is a widely used social media platform in Country A that hosts millions of users from various regions of the world. This platform facilitates users to establish profiles and connections, as well as to create online communities. Users can share various forms of content, including images and videos, and the platform offers several features such as groups catering to individuals with shared interests. The primary focus of this study is on groups dedicated to identifying employment opportunities in country A. Organisation B concentrates on this platform because of the high prevalence of suspicious posts, the ease of investigating cases, and the accessibility of essential information for their investigations, such as victim

and suspected trafficker contact details. The following section provides additional details about the job advertisement groups.

#### Job Advertisements Groups and Posts

Human traffickers frequently use job advertisement groups on Platform A to target and recruit victims. These groups are commonly used by people seeking employment opportunities, with membership sizes ranging from a few hundreds to hundreds of thousands. The majority of these groups do not have membership requirements for joining, which allows individuals from diverse age groups and backgrounds to join. These job advertisement groups can be private or public and offer interactive features that allow members to post content and comment on other posts. Joining private groups is a straightforward process that requires only an account on the platform.

The job advertisements in these groups cover a wide range of positions. The most frequently advertised roles were those of domestic workers, housekeepers, security guards, and cleaners. Individuals predominantly advertise positions on behalf of organisations or for their personal residences. The majority of job advertisements provide a written description of the position, while others present visual representations accompanied by a description of the role or include a brief message with hyperlinks to an external website containing full job advertisement details.

However, these groups are not only used for job posts; in some cases, individuals use them to request employment, seek interview tips, or find companions to help distribute their resumes within their communities.

### 5.2.2 Online Observation: Social Media Groups

I conducted an online observational study of 14 groups on Platform A for two months, with an average of 12 hours of active engagement per day. The goal was to understand first-hand the process of identifying fraudulent job advertisements and discovering patterns in these fraudulent job offers. The groups analysed during this process were identified and provided by Organisation B because they have

been found to often disseminate content related to human trafficking. The size of these groups ranged from 1,500 to 143,000 members.

Before data collection, I received training from the project manager of Organisation B. The training focused on learning the indicators used to identify suspicious job advertisements, which are discussed in Section 5.3.4. Subsequently, the observation process commenced, which involved an analysis of content from individuals posting job advertisements and those seeking employment opportunities. This process involved manual scanning of the job advertisements in each group. For each group, the posts were initially arranged in chronological order, after which job advertisements and their corresponding comments were scrutinised for indicators of human trafficking.

There was an agreement with Organisation B to forward any suspicious content discovered during the monitoring process. Organisation B viewed this as an opportunity to collect suspicious jobs that they could investigate. The reason for focussing on recent posts was to allow the organisation to assess and respond quickly if the content was found to be connected with human trafficking. After the observation period, 40 suspicious posts were reported to the project manager.

During observation, there was no direct interaction with participants in the groups, and the role of passive observer was adopted. Passive observations involve observing a group of individuals without interacting with them and are generally conducted without their knowledge (Williams, 2022). This hands-on experience helped the researchers gain a deeper understanding of the dynamics of these online groups and shed light on suspicious job postings in Country A.

### 5.2.3 In-person Observation

The objective of the in-person observations was to better understand Organisation B's method of shortlisting fraudulent job posts on Platform A and the investigation process initiated to determine whether a job is related to human trafficking. Research has shown that integrating both in-person and online observations yields a more comprehensive approach to understanding research problems (Murthy, 2008).

I carried out observations in Country A for eight days within a two-week time frame, with a daily time commitment of eight hours. During the two weeks, five locations were visited from which they normally perform their work. To select which individuals to observe, purposive sampling was used. Purposive sampling was carried out in collaboration with the project manager, and they assisted in selecting individuals who were best suited to provide the essential data needed to address the research questions. These observations followed an unstructured approach; there were no predefined behaviours at the beginning of the study. In addition, direct interaction with the participants was conducted to gain a deeper insight into the investigation procedures. During the two weeks, five investigators and one manager were observed. These observations were made during regular working hours, beginning at 8:00 or 8:30 a.m. and ending around 4:00 or 5:00 p.m.

I accompanied the investigators and managers during their efforts to raise awareness, which lasted approximately four hours each day, and then observed them while identifying and investigating potential trafficking cases on social media for an additional four hours. Observing their investigative procedures on Platform A involved listening to phone calls made to suspects and potential victims. Notes were taken to record daily events and significant findings during the observation period. In addition, discussions with the participants were recorded using audio devices and pictures of the environment were taken for the benefit of my Ph.D. supervisors. A total of 32 hours were dedicated to working closely with the teams investigating online recruitment on Platform A while 32 hours were allocated to observe other activities carried out by Organisation B.

### 5.2.4 Semi-structured Interviews

During the in-person observations, semi-structured (Barclay, 2018) interviews were conducted to complement the insights from the direct observations. These interviews were instrumental in gaining a deeper understanding of the investigation process. The questions mainly focused on understanding the reasoning behind the actions taken by the investigators and the manager when investigating suspicious jobs, as

well as clarifying any misunderstandings from the observed data. Interviews were conducted with the project manager, three investigators and the director.

The selection of interviewees was strategically determined: the project manager and director were selected on the basis of their extensive expertise in the organisation's domain and human trafficking. The investigators, on the other hand, were selected based on their availability and the guidance offered by the project manager. The interviews lasted an average of 30 minutes. The interviews were recorded and subsequently transcribed.

### 5.2.5 Case Notes

Upon investigating the cases (job advertisements), Organisation B records all cases associated with human trafficking within a dedicated software system. Before entering the cases into the software, each investigator compiles case notes on investigations that have yielded a high risk of human trafficking. These case notes cover information on the job advertisement and the investigative process. These case notes were analysed to help discover new patterns that had not previously been observed.

Organisation B granted me access to the cases over a four-month period, and 15 cases within the scope of the study were analysed. This facilitated a more comprehensive understanding of the investigation process and the indicators of human trafficking.

### 5.2.6 Analysis

Data collected from audio recordings of observations and interviews were transcribed and entered into NVivo 12 for analysis. An inductive approach (Thomas, 2006) was used to develop codes by examining study observational notes, case notes, and transcripts of the observation and interview recordings. The coding process focused on Organisation B's operational methods, indicators of suspicious job posts, understanding of traffickers' use of Platform A, investigative techniques employed by Organisation B, and potential indicators for identifying fraudulent

job advertisements. NVivo 12 Pro was used for all coding procedures and to draw patterns and conclusions.

The images of the job advertisements from the case notes and the screenshots of the job advertisements taken during the observation periods were analysed using content analysis (Bengtsson, 2016). This helped provide contextual information on the findings. Content analysis is used to analyse a variety of information, from text, visuals, and audio data. It helps identify patterns and themes in the data (Bengtsson, 2016; Erlingsson and Brysiewicz, 2017).

The combination of inductive coding, transcription, and content analysis provided a framework for examining the data and drawing meaningful insights.

### 5.2.7 Ethical Considerations of Digital Observations

Ethical approval was obtained before the start of the study under the reference CS_C1A_021_034_003. The ethical details for this chapter have been omitted to maintain the confidentiality of Organisation B's information. For observations on Platform A, a commitment was made to refrain from engaging with individuals within job advertisement groups. Furthermore, the ethical approval process clarified that a covert approach would be used and that deception would not be used when conducting online observations. Deception involves the deliberate concealment of the true objectives of a study from participants and misleadingly presenting the research, while covert research refers to the non-disclosure of research information to the participants (Spicker, 2011).

Most digital observations are generally made covertly without the knowledge of the user (Murthy, 2008). Covert observation, as part of digital ethnography, presents a range of ethical challenges, especially when covert methods are used and participants are unaware that they are being observed. Research indicates that during covert studies, researchers should ensure that participants' rights, well-being, safety, privacy, and confidentiality are respected throughout the research process A. Thompson et al., 2020). Previous studies have emphasised that researchers conducting covert research should carefully evaluate their research for several

factors before determining whether it is an appropriate method. Marzano provided a list of factors to consider, and human trafficking recruitment fell under some of these factors (Marzano, 2021).

The perspectives on covert research are frequently polarised, and some studies (Spicker, 2011) have investigated the prevalent misconception between covert research and deception. Several studies (Spicker, 2011; A. Thompson et al., 2020; Marzano, 2021) have addressed concerns associated with covert research. The authors state that before using covert research, researchers must evaluate the benefits and drawbacks of their projects, particularly when covertly observing vulnerable individuals, and make informed decisions accordingly. The consideration of these consequences is essential when determining whether to conduct research and how to present the results. This thesis assessed the benefits and potential risks of this research in Chapter 3. The following sections discuss the findings derived from the use of these methods.

## 5.3 Findings

### 5.3.1 Methods Traffickers Employ to Recruit Victims on Social Media

Human traffickers frequently employ two primary strategies to recruit victims from job advertisement groups on Platform A. The first strategy involves traffickers posting fraudulent job advertisements to attract potential victims. Victims indicate interest by publicly commenting on posts with their contact information or by contacting the traffickers using the private messaging feature on Platform A. If the victim expresses interest publicly on the platform, the traffickers engage in further communication through private messaging or by contacting them on WhatsApp. These job posts often elicited rapid responses, with more than 200 comments appearing within a few minutes after posting. These job advertisement posts often exhibit several indicators of human trafficking that raise suspicion. Organisation B expressed concern that the majority of jobs reviewed within these groups are not legitimate.

The second strategy involves traffickers analysing the posts of potential victims within the job advertisement groups to identify them. These posts frequently expressed individuals' desires for employment, ranging from casual enquiries to urgent requests. Most of these urgent requests originated from individuals who had been seeking employment for an extended period of time. Traffickers exploit this vulnerability by providing their contact information, typically WhatsApp numbers, or instructing people to contact them directly on Platform A. In some cases, potential victims include their contact details in their posts, allowing traffickers to reach them easily. Individuals in these groups frequently shared their contact information publicly.

There is also a tendency for people within these groups to advertise money-making schemes that promise easy and quick financial gains, which can lead to financial losses for potential victims. Some of these schemes share characteristics similar to advertisements associated with human trafficking, making it difficult to differentiate between them. Posts requesting companions to distribute their resumes also attract many responses.

As stated previously, Organisation B's activities can be categorised into two primary tasks: raising awareness and investigating fraudulent jobs. The following sections discuss their work further.

### 5.3.2 Organisation Work: Raising Awareness

Organisation B engages with people on the street to educate them about the dangers of human trafficking, how individuals can be recruited, and signs of human trafficking recruitment. The investigators and the project manager approach people, initiating conversations concerning the risks associated with human trafficking, while also distributing informational pamphlets. These pamphlets contain detailed information on the various methods that traffickers employ to recruit victims, along with contact details for individuals to reach out to the organisation and verify any job opportunities. The project manager speaks on these pamphlets and states

*"I look at it as an icebreaker because, on the street, that is the only way to stop strangers. Yeah, to talk to about human trafficking".*

I participated in the distribution of the pamphlets on the streets as a means of building a good rapport with the team. Language barriers were occasionally observed to pose challenges, as some people were not fluent in English. The investigators engaged with people in their local languages to overcome this hurdle, ensuring effective communication. The team encountered various reactions from the public; some people were receptive, particularly when the topic of human trafficking was raised, while others showed disinterest. The investigators noted that the older generation was more open to conversations, while the younger generation often exhibited more significant scepticism. This scepticism further highlights the crucial need to raise awareness among young people as discussed in Chapter 4, as they are a primary target of traffickers in the country, as reported by the investigators.

During these street interactions, there were instances in which individuals shared their personal experiences with trafficking or disclosed suspicious opportunities that they were currently facing. For example, during the observation period, an individual informed the team that their sister had attended a job interview and was later found dead. Such stories stress the significance of the work conducted by Organisation B. The investigators mentioned that they often come across such stories during their work.

As part of their street engagement, the investigators also seek commitments from the people; they ask people to sign the pamphlets as a commitment to disseminate information about trafficking within their communities. The investigators capture photographs of these commitments. These images are shared in a group chat with the project manager and the director as tangible evidence of their work.

Furthermore, during these awareness campaigns, the staff communicates to individuals that they have the option to contact the Organisation B to verify employment opportunities. The pamphlets include instructions on how to reach out to them to verify these job advertisements.

The investigators who distribute the pamphlets include their names on them. When Organisation B receives a message from a prospective victim to verify opportunities, the project manager identifies the investigator who initially spoke with the individual. Subsequently, it becomes the investigator's responsibility to conduct an investigation of the opportunity in question.

The investigators also use Platform A to identify and scrutinise potential fraudulent employment prospects. The following section discusses the process of identifying and examining job opportunities on Platform A.

### 5.3.3 Organisation Work: Tackling Online Human Trafficking Recruitment

The work of Organisation B in identifying and investigating suspected cases of human trafficking on Platform A can be summarised into four primary stages, as illustrated in Figure 5.2.

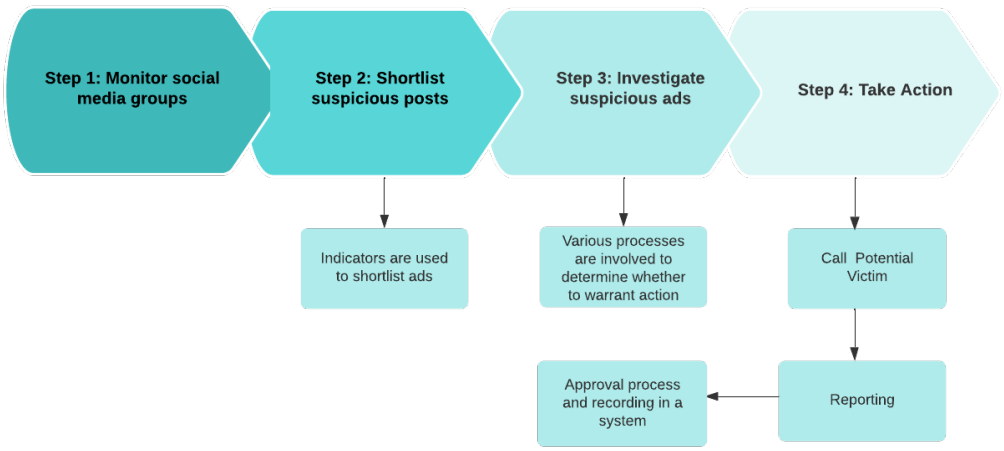


**Figure 5.2:** Overview of Organisation B's Workflow

The first stage involves monitoring Platform A and identifying potentially suspicious job advertisements. Subsequently, the organisation conducts an investigation to determine whether these job advertisements are associated with human trafficking. If sufficient evidence is obtained, the organisation informs potential victims of the risks inherent in the job opportunity and advises against pursuing it.

Organisation B manages these cases independently and does not report them to the authorities or collaborate with law enforcement. Instead, they communicate directly with the at-risk individuals. This lack of collaboration is further discussed in Section 5.4. The following sections provide additional details of these four stages.

### Monitoring and Shortlisting Advertisements

The monitoring stage involves examining social media groups to identify job advertisements that may be linked to human trafficking. As mentioned previously, Organisation B has a set of indicators to determine the legitimacy of job advertisements in Country A. Investigators receive regular training to stay up-to-date on the evolving tactics used in human trafficking.

The investigators review job advertisement posts from the predetermined list of 14 groups provided by the Project Manager during online observations, while also exploring other job advertisement groups within Country A on the platform. In addition, they employ keyword searches to identify suspicious job advertisements and subsequently monitor the associated groups or pages. Investigators do not rely solely on a single indicator when selecting suspicious posts; instead, they base their judgement on a combination of indicators. Most investigators also use a unique method to assess content on Platform A because they have developed their own set of indicators.

These identification processes are conducted manually, without technological assistance. The project manager occasionally supports the monitoring process; however, investigators are primarily responsible for identifying and investigating these cases.

### Investigation Process

Investigators typically conduct their analyses independently; however, they occasionally seek consultation from colleagues when confronted with the ambiguous aspects of a case. They frequently consult with the project manager 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** - Verifying the legitimacy of a company mentioned in suspicious job advertisements is a crucial initial step. This process involves cross-referencing company information with an online public system that maintains records of registered companies, including their contact details, locations, and essential personnel information. By comparing the details provided in the job post with the data available in the system, Organisation B can ascertain whether the company is legitimate. If the investigation reveals that a company does not exist or is no longer operational, it raises significant concerns that require further examination. Conversely, if the company is confirmed to be legitimate, the analysis compares the company's particulars with those disclosed in the job advertisements. Any discrepancies or inconsistencies discovered between job posts and organisational information raise suspicion of trafficking.

2. **Investigate the phone number of the suspected trafficker** - Organisation B also uses a third-party application to verify the validity of the telephone number associated with the suspected trafficker. This third-party application provides information on individuals connected to the phone number, including their name, registration status, and whether the number is classified as spam. This application is publicly accessible and can be used by all individuals. If there is a discrepancy between the name associated with the phone number and the information provided in the job advertisement, it raises concerns about the legitimacy of the job post.

3. **Investigate the role and packages offered** - Organisation B also examines the incentives for each job post to verify its authenticity. Using multiple resources to determine the typical compensation package for the advertised position, the organisation can accurately assess whether the proposed remuneration aligns with the prevailing industry standards. The Project Manager discusses the need to understand industry packages:

> "One of the things that we encourage staff is to understand what is going on in the country and understand the different industries and the packages, the requirements and stuff" ( Project Manager).

4. **Call the suspected trafficker** - Organisation B also initiates contact with suspected traffickers using the telephone numbers provided in the job advertisement, assuming the role of a prospective candidate to elicit information. During these interactions, they observe indicators of deceptive behaviour, false statements, or exaggerated claims made by suspected traffickers. Investigators often employ deceptive enquiries to assess the credibility of suspected traffickers, such as enquiring whether the absence of employment records or an unlawful presence in the country impedes their ability to pursue the opportunity. Throughout the observation period, I noted that each of the suspected traffickers contacted asserted that the investigators' lack of legal authorisation to work in the country would not present an impediment.

5. **Call potential victims** - Organisation B obtains additional information by contacting potential victims and soliciting any further details they may possess regarding the employment opportunity. This communication often results in victims divulging supplementary information, such as the precise details about the interview such as time and location. Organisation B proceeds to review the additional information.

6. **Reviewing job application links** - Organisation B examines the job application links incorporated within suspicious job advertisements. These links often result in circular navigation patterns or provide only limited information.

### Taking Action and Reporting

Upon accumulating substantial evidence indicative of potential fraudulent activity, Organisation B initiates a series of procedural actions.

1. **Contact the potential victim** - Investigators first initiate communication with potential victims using a phone call, where they express their concerns about the job opportunity and provide rationales for their suspicions. In addition, they warn against further engagement with the advertised position. Organisation B subsequently refrains from any additional communication with the victim following this initial contact.

2. **Write up the case** - Investigators document all relevant information related to the case, including a copy of the job advertisement, any interactions with the suspected trafficker, and the information used to determine that the job advertisement was related to human trafficking. This case report is essential for subsequent evaluation and authorisation. At this juncture, the case is classified as a 'high risk of human trafficking'. The investigators also complete customised versions of Form 1 and Form 2 discussed in Chapter 4, along with three additional forms that record information on potential victims, suspected traffickers, and locations.

3. **Approval process** - Upon completion of the writing process, a data analyst assesses whether all the requisite documents pertaining to the case have been provided by the investigators and subsequently proceeds to upload them to a software system. Once the case has been uploaded, two senior staff members, including the project manager, examine all the relevant details of the case and independently formulate a recommendation regarding whether to classify it as having "evidence of human trafficking". As previously discussed in Chapter 4, this classification indicates that the case has sufficient evidence of human trafficking.

   In instances where the two individuals do not reach a unanimous decision regarding their assessment of the case, a third senior member of the staff is consulted to make a definitive judgement. In the majority of cases, the disagreement between the two members typically involves whether the case should be considered high-risk or whether it has sufficient evidence of human

trafficking. If a case is deemed to possess sufficient evidence, it is officially documented as "evidence of human trafficking"; otherwise, further investigation is necessary.

4. **Record storage and access** - Following the approval of a case as "evidence of human trafficking", modifications are no longer permissible. Consequently, the case is documented and archived on a digital platform used by Organisation B for the monitoring and management of all cases under its purview.

Organisation B only examines cases in which contact information for both the trafficker and victim is available. Communication with suspected traffickers and victims constitutes a fundamental aspect of their investigation. In addition, Organisation B's capacity for intervention is limited to dissuading the victim from pursuing the opportunity as they do not work with law enforcement agencies. These contact details are often taken from the job advertisement and corresponding comments section.

### 5.3.4 Identifying Suspicious Job Advertisements

This section discusses the indicators of job advertisements used to determine whether they are associated with human trafficking. The educational background and experiences of the victims may, in some cases, impede their capacity to recognise certain indicators outlined in this section.

The job advertisement screenshots were modified through the removal of timestamps and other identifying graphics using cropping tools to mantain annonymity of Platform A. These indicators are categorised into five primary themes. A summary of these themes and their corresponding indicators is presented in Table 5.1.

#### 1. Attract Wide Audiences

A common strategy used in fraudulent job advertisements associated with human trafficking is to attract a large number of applicants, thus increasing the likelihood of successful recruitment. This is typically accomplished through four primary

| Theme | Indicators |
|---|---|
| 1. Attract Wide Audiences | Company Names |
| | Recruitment Volume |
| | Multiple Roles |
| | Preferences |
| 2. Use of Informal Writing | Writing Styles |
| | Emojis |
| | Suspicious Links |
| 3. Offer Attractive Packages | Salary |
| | Incentives |
| 4. Use of Known Trafficking Details | Similar Job Advertisement Details |
| | Trafficking Locations |
| 5. Request for Personal Information | Personal Details |

**Table 5.1:** Themes and Indicators

methods: the use of *well-known company names*, the advertisement of *multiple roles and mass recruitment*, and the inclusion of *specific preferences*.

**a. Company Names -** Concerns arise when a job post for a position at a *reputable company* appears in job advertisement groups on Platform A. Typically, established companies employ more conventional recruitment methods rather than social media platforms especially on these types of job advertisement groups.

Human traffickers frequently select well-known brands, particularly in the food sector, to advertise these job advertisements. Organisation B has repeatedly observed and confirmed that recognised companies rarely use Platform A for their staffing requirements.

**b. Recruitment Volume -** Job postings from human traffickers often claim to seek a *substantial number of applicants*, sometimes in the hundreds. This tactic raises suspicions of potential scams, as it is uncommon for legitimate companies to advertise such a large number of job openings in a single posting. An example of this type of job advertisement, which seeks numerous applicants, is shown in Figure 5.3.

These advertisements also typically deviate from the professional tone and structure used by genuine organisations in their recruitment materials. Organisation

B successfully intervened to prevent a potential victim from accepting a similar offer, which could have led to exploitation.

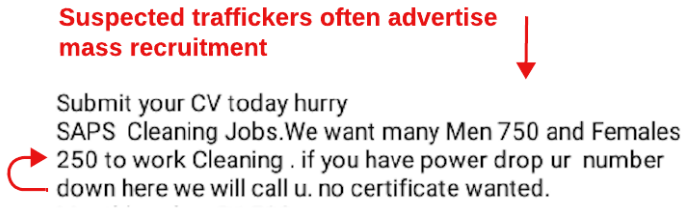


**Figure 5.3:** Indicator Showing Mass Recruitment

**c. Multiple Roles -** Job posts that raise suspicion often contain *multiple positions* in a single advertisement, potentially aiming to attract a diverse range of applicants by offering various employment prospects. These posts merit careful examination, particularly when they feature more than three roles, as they frequently deviate from the typical structure of genuine job advertisements. Moreover, the positions promoted in these listings are commonly associated with human trafficking recruitment. Such roles include security personnel, general labourers, and cleaning staff. An example of this indicator is shown in Figure 5.4.




**Figure 5.4:** Example of a Job Advertisement - Two Posts Advertising Multiple Roles

**d. Preferences -** Human taffickers also often post job advertisements that express *particular preferences* such as demographic requirements, which are normally used by traffickers to broaden their scope and target more vulnerable individuals. For

example, traffickers might reduce or eliminate job qualification criteria to appeal to those lacking experience or specific credentials, thereby enlarging their pool of potential victims. Furthermore, they target vulnerable groups by specifying countries with a high number of undocumented immigrants who are often desperate for employment opportunities. Such advertisements are beneficial for traffickers as they lure many potential victims with the prospect of work. Furthermore, illegal immigrants are also preferred by traffickers because they can be easily controlled due to their fear of being reported to law enforcement (United Nations Office on Drugs and Crime, 2020).

The job post depicted in Figure 5.5 stated that no prior experience was necessary for the application, encouraging potential victims to express interest and contact the trafficker. Fortunately, Organisation B was able to confirm its fraudulent nature and prevent the recruitment of one victim. The phrase "No experience required" is frequently used and can be searched on Platform A to identify suspicious job advertisements. The *writing style* of these advertisements also serves as a crucial indicator, which is elaborated on in the following section.

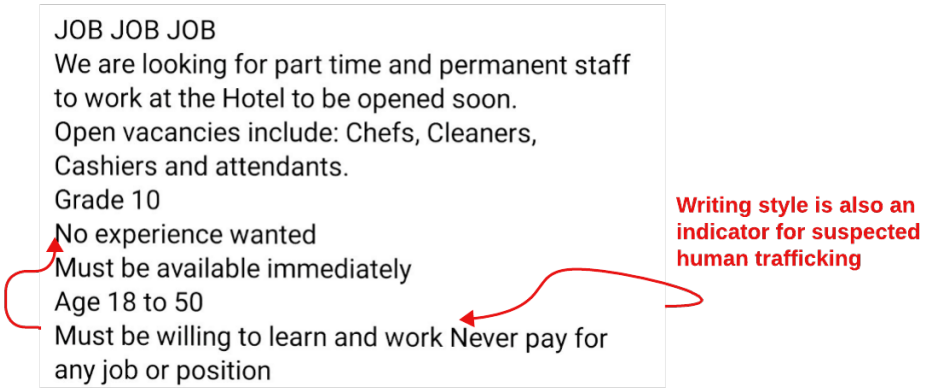


**Figure 5.5:** Indicator Showing Writing Style

### 2. Use of Informal Writing Styles

Job advertisement posts associated with human trafficking also typically employ an informal tone, often incorporating *informal writing*, *emojis*, and *suspicious*

*hyperlinks*. These elements are generally not observed in legitimate job advertisement posts.

**a. Informal Writing -** The *writing style* employed in job advertisements plays a crucial role in identifying suspicious posts. Human traffickers frequently create job posts that contain grammatical and spelling errors as well as informal language, in contrast to the formal tone typically observed in legitimate job advertisements. For example, as shown in Figure 5.5, these advertisements contain grammatical errors and elicit concerns.

In addition, traffickers aim to attract people to consider fraudulent opportunities by using language (slang) that resonates with vulnerable people desperately seeking employment.

**b. Emojis -** Incorporating *emojis* into job listings is an uncommon and often questionable practice. Human traffickers use various emojis, such as currency symbols or national flags, to attract potential victims and entice them to submit job applications. In developing countries, job seekers typically explore employment prospects in more developed countries. Consequently, emojis that represent money or specific countries can easily grab their attention and motivate them to pursue such opportunities. An illustration of this tactic is presented in Figure 5.6, which shows a job advertisement featuring Canadian flags.

**c. Suspicious Links -** Human traffickers frequently incorporate *misleading hyperlinks* within their advertisements, which serve as an additional indicator of fraudulent job posts. These hyperlinks typically redirect users to fabricated job pages rather than pages with genuine employment opportunities. The presence of such deceptive link patterns is considered in assessing whether a given advertisement is associated with human trafficking.

### 3. Offer Attractive Packages

Human traffickers also commonly employ unrealistic job packages to attract and engage potential victims. This section examines *salaries* and various *incentives* as potential indicators for suspicious job advertisements.

**a. Salary -** Traffickers frequently resort to offering unusually *high salaries* as an incentive to recruit victims. Such salaries are often considerably higher than average remuneration for comparable positions in Country A. Advertisements that promote employment opportunities, with salaries exceeding twice the average amount, are particularly suspicious. Experts familiar with the prevailing salaries and remuneration packages in a given country can easily identify these unrealistic offers.

**b. Incentives -** Traffickers also often employ recruitment strategies that include highly attractive incentives, such as complementary transportation and accommodation, to entice individuals to respond to their advertisements. Organisation B has encountered multiple cases in which individuals were offered free transportation for interviews. By using these appealing offers, traffickers can effectively capture the attention of potential victims. In Figure 5.6, a trafficker pledged to cover the expenses of visas and flights for potential victims, prompting several individuals to initiate contact with suspected traffickers. Organisation B indicated that traffickers often assume these expenses when transporting victims to another country, subsequently using them to establish control over victims through debt bondage if the victims are unable to reimburse the costs. This was also evident from the analysis in Chapter 4.

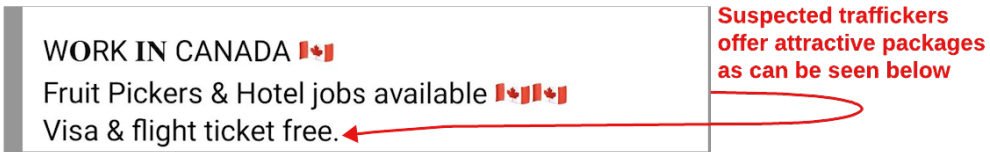


**Figure 5.6:** Example of Job Advertisements - Indicator Offering Attractive Packages

### 4. Use of Known Trafficking Details

Suspicious job posts can also be identified by recognising specific details associated with known offenders in previous cases. This section examines the significance of *similar advertisements* and *well-known locations* as crucial indicators.

**a. Similar Job Advertisements -** Traffickers frequently employ the practice of*duplicating information* when creating deceptive advertisements. Organisation B can identify such advertisements by referencing previous investigations. However, investigators and project managers can easily recall details associated with suspicious job advertisements that frequently appear in these groups, whether from distinct accounts or the same accounts. Suspected traffickers typically minimally modify job posts 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. The Project Manager discusses this issue of duplicating information:

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

**b. Trafficking Locations -** The presence of *known trafficking locations* also plays a significant role in identifying potentially suspicious job advertisements. Organisation B has developed a list of locations linked to human trafficking operations within Country A. Job advertisements that mention these locations are subsequently flagged by the organisation for additional investigation. At times, group members also make cautionary remarks on job posts with suspicious locations to warn potential victims of exploitative schemes.

### 5. Request for Personal Information

Job posts that solicit personal information in comments require scrutiny, as they can be exploited by human traffickers to obtain the contact details of potential victims. Traffickers often use these details to communicate with potential victims

about interview invitations outside the comment section. People in these job advertisement groups often express scepticism concerning job offers within groups, as they receive interview invitations from companies to which they have not applied, thus raising concerns.

### 5.3.5 Characteristics of Suspicious profiles

This study observed patterns in profiles related to potential human traffickers, which can be used to identify them and the job advertisements that they post. These accounts show distinctive features that can be detected and flagged.

1. Account Profile - Individuals who disseminated suspicious advertisements on Platform A typically had profiles that exclusively contained employment-related content. On Platform A, users can view an account holder's previous posts if their profile is public. Most of the content of suspected traffickers consisted of job posts in various industries and companies, many of which exhibited suspicious characteristics.

2. Account Photo - The profile images associated with suspected traffickers also elicit concern. Suspected traffickers either frequently lacked a profile picture or displayed an image that was not of themselves. These profiles often featured quotations or other non-personal elements. In some instances, the accounts used photographs of renowned models as their profile picture.

3. Connections - The accounts under examination generally showed minimal or no connections, raising suspicion. The primary activity on these accounts involved sharing employment opportunities, as mentioned above, and, in some cases, promoting money-making schemes.

## 5.4 Discussion

### 5.4.1 Summary

This chapter examines a collaborative effort with Organisation B to understand the strategies traffickers use to recruit individuals through social media platforms, while

also highlighting the organisation's operational procedures. Particular attention is paid to the investigative techniques used by experts to identify fraudulent job advertisements online. The following sections analyse the organisation's work, explore opportunities for using technology to combat online recruitment, and consider the design implications for social media platforms in preventing human trafficking.

### 5.4.2 Analysis of Organisational Work

Organisation B employs a method that separates it from similar entities in other countries in which Organisation A operates. This distinctive approach stems from the unique work of the organisation, which yields several significant advantages in combating trafficking while simultaneously posing certain difficulties.

**Benefits**

Benefits observed include:

1. **Early Intervention**: Organisation B's approach allows for early intervention by reaching out to potential victims at the initial stages of recruitment. This proactive attitude helps prevent people from suffering trauma and incurring debts associated with trafficking, such as travel expenses, accommodation costs, and visa fees. Early intervention disrupts the recruitment process, safeguarding individuals from further exploitation (Volodko et al., 2020).

2. **Community Empowerment**: Organisation B's awareness campaigns and pamphlet distribution enable individuals to share information within their communities. This approach raises awareness of the dangers of human trafficking and equips people with the knowledge to identify and report possible cases. The ripple effect of this empowerment extends beyond Country A, as individuals who receive pamphlets often share them with their family and friends, spreading the message of vigilance and prevention. Organisation B has seen its pamphlets being shared online and in other countries.

3. **Active Investigation on Platform A:** Organisation B actively investigates fraudulent job advertisements on Platform A, recognising the platform as a significant hub for human trafficking recruitment. Given the high rates of youth unemployment in the country, many young people are vulnerable to exploitation and recruitment. By intercepting and exposing these fraudulent job postings, the organisation prevents potential victims from being successfully recruited.

**Challenges**

The following are the challenges and concerns that were identified:

1. **Lack of Law Enforcement Collaboration -** In contrast to other countries, Organisation B does not engage in direct collaboration with law enforcement agencies. This lack of cooperation may have adverse consequences for personnel and potential victims. Staff members may be exposed to hazardous situations when operating independently, which can impede access to crucial information or restrict their capacity to perform duties. For potential victims, human traffickers can persist in their illicit activities and continue to exploit new individuals, perpetuating the human trafficking cycle. Attempts to partner with law enforcement agencies on this work have proved challenging for Organisation B, with no clear reasons given by the agencies for the lack of collaboration.

   Previous research indicates a general lack of collaboration in the fight against human trafficking between governments and anti-trafficking organisations in Africa (Onuoha, 2011). This insufficient collaboration can be attributed to various factors that vary between countries. Governments' operational methods can influence the approach adopted by law enforcement to address human trafficking. As mentioned in Chapter 4, governments in Africa tend to prioritise prosecution in their anti-trafficking strategies (Britton and Dean, 2014), with law enforcement focussing predominantly on investigating potential trafficking cases, identifying and rescuing victims, and apprehending traffickers

rather than prevention. In contrast, most anti-trafficking organisations adopt a victim-centred approach, they strive to prevent human trafficking, identify victims, and provide support to survivors. The difference in objectives between governmental and non-governmental stakeholders can contribute to the overall lack of collaboration.

The need for law enforcement and anti-trafficking organisations to maintain the confidentiality of their data and operations can also prevent their collaborative efforts. Stakeholders frequently operate under stringent data processing regulations that prohibit the sharing of certain information. Consequently, there is a requirement to develop methods that facilitate cooperation between stakeholders without compromising privacy. Certain countries, such as Namibia and Kenya, where the organisation also operates, have established successful collaborations between anti-trafficking organisations and governments, including law enforcement agencies. More research is required to understand the processes involved in the formation and maintenance of these partnerships in such countries. The absence of collaboration between law enforcement and anti-trafficking organisations is not exclusive to Africa, but is also evident in Europe (Campana, 2022), suggesting that it may constitute a global phenomenon.

2. **Staff Safety Concerns -** The elevated crime rate in Country A presents substantial risks to Organisation B's personnel. Various incidents have occurred during investigative work, such as when intoxicated individuals attempted to harm an investigator and when investigators witnessed a robbery in the communities where they operate. This violence causes concern for staff and can also impede Organisation B's ability to intercept potential victims.

3. **Few Jobs are investigated -** Organisation B's manual method for collecting and analysing advertisements presents a substantial challenge in addressing the high volume of fraudulent job postings. Manual examination of content in groups comprising thousands of job posts and scrutinising each suspicious

advertisement is a labour-intensive process, which limits the number of job posts that can be reviewed by investigators, resulting in numerous fraudulent posts remaining undetected.

In addition, certain job advertisements are not subjected to investigation because of Organisation B's requirement of obtaining both the trafficker's and victim's contact information to initiate an investigation. Traffickers and victims do not always disclose their contact details; they occasionally opt to communicate privately through messaging services on Platform A.

Organisation B also faces constraints in financial and non-financial resources, which impedes its capacity to address a substantial volume of fraudulent job postings. Consequently, many suspicious posts remain unresolved. Most anti-trafficking organisations have limited resources and may also operate manually, like Organisation B, causing similar problems.

The following section explores how technology can be leveraged to assist Organisation B in identifying and investigating cases more efficiently.

### 5.4.3 Opportunities for the Use of Technology to Identify and Investigate Suspicious Job Posts

The findings in this chapter highlight the importance of using technology to assist stakeholders in combating false job postings used for human trafficking recruitment. This section examines the various ways in which technology can be applied to address this issue and help Organisation B and other organisations with their work operations.

#### Reviewing Job Posts Online

The findings of this study confirm the need for technology to identify suspicious posts related to human trafficking. Automating the process of identifying suspicious posts can significantly enhance Organisation B's efficiency in managing a greater number of job advertisements while also enabling the organisation to reallocate resources to other critical tasks. However, as previously noted, there is a lack

of tools that target the recruitment stage of human trafficking despite evidence indicating that the majority of recruitment activities occur online in several countries (Progetto Tenda et al., 2021; Feehs and Wheeler, 2020; United Nations Office on Drugs and Crime, 2020)..

Designing anti-trafficking software for the recruitment stage can benefit from incorporating advanced technologies that understand natural language to analyse and classify suspicious job postings. LLMs have been shown to outperform traditional machine learning approaches and excel in classifying social media posts (Chae and Davidson, 2023). Other methods that excel in understanding context can also be reviewed.

Anti-trafficking tools analysing job posts should categorise the advertisements with varying levels of certainty based on a defined scale. This approach will help stakeholders reviewing the jobs to determine appropriate procedures and follow-up steps for each level.

The indicators identified in this study can help review job postings online. However, the widespread use and availability of LLMs may soon render some indicators irrelevant, such as the use of informal writing style. Traffickers can leverage LLMs to create more realistic job advertisements, making detection more challenging. Therefore, efforts should always be made to assess the viability of these indicators during the design and development stage.

### Investigating Job Posts

There is an opportunity to develop a tool that enables individuals to seek verification of opportunities from Organisation B, making it easier to review and manage cases, deviating from the current approach of reaching out to Organisation B using a messaging application. This software tool could include features that automate the investigation process, 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. Additionally, a location checker could automatically assess whether the specified location in an advertisement is associated with well-known trafficking hotspots. This software tool would help Organisation B to be more efficient in performing its duties and investigating more cases.

**Identifying Potential Traffickers and Organised Networks**

Previous research has also shown that trafficking networks operate on various online platforms, exploiting multiple victims (Arenas et al., 2015; Ardakani, 2020). The findings in this chapter also reveal patterns in the profiles of suspected traffickers. There are opportunities to use different technologies to flag potential traffickers and organised trafficking networks on social media. These networks can be identified on the basis of their connections and interactions.

For example, a previous study (Ardakani, 2020) identified organised groups of sex traffickers in Louisiana by analysing online advertisements. By using the advertisement's ID (primary link) and phone numbers, they connected different advertisements to create networks. In addition, they applied author attribution techniques to the textual content of the advertisements as a secondary link, helping to identify advertisements posted by the same entity. Similarly, another study (Arenas et al., 2015) identified trafficker networks on Twitter that promoted OnlyFans accounts of their victims.

As these studies demonstrate, social network analysis is effective in uncovering trafficking networks. In addition, technologies such as AI models can be trained to identify suspicious profiles based on the characteristics outlined in Section 5.3.5 and Chapter 4, further enhancing efforts to detect trafficking activity.

Limited research has used content generated during the recruitment stage to identify human traffickers and their networks. Identifying human traffickers and their organised networks during this stage can assist platforms in flagging traffickers and prevent them from disseminating advertising opportunities, thus reducing the recruitment of individuals. Furthermore, there exists an opportunity to verify whether organised behaviour can be identified through mentions and links in job

posts, similar to the approach adopted by previous studies (Arenas et al., 2015) in analysing the content of OnlyFans posts.

Employing methods to identify traffickers and organised crime would also be beneficial in cases where online content is insufficient to determine whether it is associated with human trafficking. Reports show that most of the cases prosecuted involve traffickers who are part of organised groups (United Nations: Office on Drugs and Crime, 2023). Identifying these groups would be highly beneficial in preventing the recruitment of victims.

### Identifying Potential Victims

Previous research emphasised the importance of identifying patterns in individuals recruited through digital means (OSCE, 2020). This knowledge will facilitate social media platforms to identify potential victims, enabling the implementation of targeted measures to protect these individuals from exploitation. Such measures may include developing features that provide pertinent information to these individuals and encourage them to verify employment offers or other opportunities with organisations such as Organisation B.

### Online Campaigns

One of the most significant issues facing social media users today is the risk of divulging personal information (Arronte, 2018; Lee et al., 2021; Caers et al., 2013) without fully understanding its potential consequences. Unfortunately, this practice is common and can lead to dangerous outcomes, particularly for individuals in work-related groups. Traditional recruitment platforms, such as job boards, usually provide a certain degree of security by protecting the contact information of potential candidates, but it is not the same for social media platforms. Therefore, it is necessary for anti-trafficking stakeholders, including governments, to take action by launching online awareness campaigns aimed at educating users about the risks of sharing personal details on social media. These online campaigns should be designed to target different audiences, especially those at risk.

### 5.4.4 Design Implications for Social Media Platforms

Social media platforms are required to protect users from harm online, including human trafficking (Department of State, 2023; ActiveFence, 2023). These organisations should recognise the potential risks that lack of monitoring can pose to vulnerable and marginalised communities, particularly if users are unaware of the dangers associated with online interactions (Department of State, 2023). This section presents a set of recommendations for social media companies to consider when designing, developing, and iterating their platforms to promote safety against human trafficking recruitment.

**Stringent Guidelines** It is recommended that designers implement stricter criteria for the entry and posting of content in job marketplaces and groups, as the current admission process poses significant risks to vulnerable users. Social media platforms should proactively establish and enforce rules for job-related groups, rather than relying solely on community moderation and group administrators.

Social media platforms have the resources and data necessary to develop effective methods for detecting and removing fake profiles from their networks, thereby preventing these profiles from accessing job marketplaces and groups. The prevalence of fraudulent profiles on social media platforms has made it easier for traffickers to identify and target victims (Arronte, 2018). During this study, it was observed that accounts that posted fraudulent jobs often exhibited fake profile attributes, as indicated by their account information and profiles. Social media organisations can use the indicators of suspicious account profiles outlined in Section 5.3.5 as an additional method to identify fraudulent profiles.

**Warning Mechanisms** Social media platforms should integrate warning mechanisms 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 job 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 behaviour in various areas, including health (Perry et al., 2015).

**Cautionary Messages** Social media platforms should implement cautionary messages to discourage users from sharing personal information when posting this information in these job advertisement groups. 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 advertisement posts 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.

**Flagging Job Posts and Accounts** Social media platforms should take a proactive approach to analysing job posts and accounts for signs of human trafficking to prevent them from posting job advertisements (Department of State, 2023). Mechanisms such as time delay can be implemented while a post is reviewed. Automated means of analysing job posts and accounts should involve experienced individuals in making final decisions before implementing drastic measures, such as removing an account from the platform.

**Reporting Mechanisms** Social media platforms should implement mechanisms by which users can report suspicious activities online (Department of State, 2023; Anderson et al., 2024). The research findings indicated that certain individuals within groups use the comment section to warn others about specific job advertisements and advise against pursuing opportunities based on information known to be

associated with human trafficking. Providing users with a means to flag such posts and provide rationales under the job advertisements could prove highly beneficial. For example, implementing a feature similar to Twitter community notes, which enables the public to append contextual information to posts (Twitter, n.d.), could alert users about potentially high-risk job posts. Anti-trafficking stakeholders such as Organisation B could also use such features to disseminate information after investigations, warning individuals about fraudulent advertisements, and extending their reach beyond those who have provided their contact details.

However, challenges can arise with the use of features like Community Notes. Since they rely on user contributions, the information provided may sometimes be inaccurate, biased, or even deliberately misleading. To mitigate these risks, social media companies could restrict the ability to add such notes to verified organisations that actively work to combat human trafficking or similar crimes, ensuring that the information shared remains credible and trustworthy.

## 5.5 Reflection on Fieldwork

The fieldwork conducted in this study presented various risks, with the primary concern being potential physical harm. Country A has been reported to have high crime rates, particularly violent attacks, which require extra precautions for foreign visitors.

Other risks included emotional distress due to the sensitivity of the research topic and health risks, such as potential disease outbreaks. In addition, some unexpected risks emerged during fieldwork, such as the risk of witnessing crimes or experiencing disturbances from drunken people while conducting awareness campaigns. These challenges required adaptability and careful management in the field with the help of the project manager.

The university follows guidance from the Foreign, Commonwealth& Development Office (FCDO) regarding researcher travel. At the time of my trip, the FCDO did not advise against travelling to Country A. In such cases, the university requires researchers to complete a risk assessment form, which must be approved by the

supervisors. The risk assessment allowed me to anticipate potential risks, including those mentioned above, and identify mitigation strategies, such as using Uber for all transportation to improve personal safety. In addition, the university provided optional training courses for my supervisors and me on various topics, such as personal safety during interviews. During my time abroad, the university offered support, primarily through check-ins with my supervisors.

## 5.6 Conclusion

This study used a multifaceted approach that included in-person and digital observations, expert interviews, and case analysis to understand human trafficking recruitment on a popular social media platform. The findings revealed the features leveraged by traffickers on Platform A, and indicators of suspicious job postings for human trafficking. The findings also discussed the strategies used by anti-trafficking experts to identify recruitment on the platform. Based on these insights, design considerations were suggested for social media platforms to prevent recruitment activities. In addition, opportunities to leverage technology to identify fraudulent jobs on online platforms were also highlighted. The findings contribute to the work of promoting safer digital environments.

The following chapter will focus on developing a labelled dataset of fraudulent job advertisements associated with human trafficking recruitment in Country A. It will subsequently employ machine learning approaches to classify fraudulent job opportunities related to human trafficking.

# 6

# Benchmarking AI Techniques to Detect Fraudulent Online Job Advertisements Related to Human Trafficking

## Contents


**6.1 Introduction . . . . . 197**
**6.2 Human Trafficking Recruitment Dataset . . . . . 199**
6.2.1 Data Collection and Preprocessing . . . . . 199
6.2.2 Data Processing . . . . . 200
6.2.3 Data Annotation and Coding Scheme . . . . . 201
6.2.4 1K Dataset Statistics . . . . . 206
**6.3 Classification: Benchmarking Various Approaches . . . 207**
6.3.1 Traditional Machine Learning Approach . . . . . 207
6.3.2 Pre-trained Neural Language Models . . . . . 215
6.3.3 Large Language Models Approach . . . . . 222
**6.4 Expert Feedback . . . . . 224**
**6.5 Trafficking Recruitment 10K Dataset . . . . . 226**
**6.6 Discussion . . . . . 227**
6.6.1 Summary of Findings . . . . . 228
6.6.2 Analysis of Classification Results . . . . . 228
6.6.3 Analysis of Job Advertisements . . . . . 231
6.6.4 Dataset Design Implications . . . . . 233
6.6.5 Future Technical Implications . . . . . 235
6.6.6 Recommendations for Policymakers . . . . . 237
**6.7 Conclusion . . . . . 238**

## 6.1 Introduction

This chapter builds on the ongoing collaboration with Organisation B in Country A to address the fourth research question: *How can a set of models be developed to identify fraudulent job advertisements for online human trafficking recruitment?* At the time of this study, I had not come across any dataset on fraudulent jobs for Country A. Therefore, the goal is to create a labelled dataset to train the models. The dataset will be limited to job advertisements from Country A, where experts have the highest level of expertise.

The primary objective of creating the dataset is to classify fraudulent job advertisements for human trafficking using traditional machine learning techniques, PLMs, and LLMs. Traditional machine learning methods offer several advantages, such as requiring fewer computational resources, achieving good results with limited data, and providing easy interpretability, which makes them ideal for anti-trafficking organisations with limited resources (V. Khanna, 2024; Explainable AI, 2024). PLMs and LLMs achieve state-of-the-art results in text classification tasks (Islam et al., 2024; Vaswani et al., 2017; Devlin et al., 2018; Garrido-Merchan et al., 2023). However, they have not been extensively tested in this field, particularly with regard to human trafficking recruitment. By comparing the performance of these three approaches in this thesis, it will be possible to determine which is the most effective in detecting fraudulent job advertisements related to human trafficking, particularly those that pose a high risk. The approach that delivers the best results will be used to label unlabelled data, which will then be assessed and validated by experts. If the results are satisfactory, the best approach will be used to label a larger number of advertisements to form a larger machine labelled dataset.

The aim was to use Platform A, discussed in Chapter 5 and build on our previous work. However, Platform A denied requests to gain access to public data in its job advertisement groups. In addition, Platform A has legal implications for accessing data on their platform using automated methods. Fortunately, Organisation B

informed us that there are two job boards commonly used by traffickers where we could acquire data, which will be referred to as Website A and Website B. Analysis of the data in Chapter 4 also showed that some victims were recruited through these websites in Country A. Website A was more commonly associated with human trafficking than Website B.

Flagging fraudulent job postings related to human trafficking will encompass all forms of trafficking, rather than focussing on a specific type, such as sex trafficking. When classifying job advertisements for human trafficking, experts have no definitive way to determine the specific type of trafficking victims are being recruited for based solely on the job postings. Furthermore, victims may be subjected to multiple forms of trafficking after recruitment, such as sex trafficking and forced labour, as discussed in Chapter 5.

This chapter makes the following contributions.

1. Presents an expert-labelled dataset on fraudulent job advertisements related to human trafficking. This dataset will be valuable not only for this thesis, but also for other stakeholders in the private and public sectors.

2. Produces a set of models to classify fraudulent job advertisements, laying the groundwork for a bespoke tool for Organisation B and other stakeholders.

3. Benchmarks the performance of large language models, pre-trained neural language models against traditional machine learning approaches in classifying fraudulent job advertisements linked to human trafficking. To the best of my knowledge, this has not been done before in this field.

4. Presents a larger machine-labelled dataset on fraudulent jobs related to human trafficking using the best-performing model.

5. Provides additional insights and patterns on online human trafficking recruitment, contributing to a deeper understanding of the issue. This new knowledge will improve our understanding of the dynamics of online recruitment in the context of human trafficking.

This chapter is organised as follows. Section 6.2 discusses the methodology used to collect and produce a 1K expert-labelled dataset. Section 6.3 outlines the process of classifying the dataset using different approaches, while Section 6.4 covers the validation of the best model with experts and the creation of a 10K machine-labelled dataset. Section 6.5 provides a summary of the analysis of the 10K dataset. Section 6.7 summarises the findings, discusses their implications, and explains how the P3D Framework guided the development of the model and datasets. Finally, Section 6.7 concludes the chapter with a discussion of the limitations of this work.

## 6.2 Human Trafficking Recruitment Dataset

### 6.2.1 Data Collection and Preprocessing

The initial strategy involved scraping the data from websites A and B using automated methods. A series of queries were formulated to collect data based on the insights gained in Chapters 4 and 5. However, the ethics review committee did not approve the use of the data, citing legal reasons and stating that the scraped data could not be used without obtaining consent from the websites involved. Despite many attempts to obtain consent from websites A and B, it was not granted. A summary of the scraping process and the steps taken to clean and pre-process the data are provided in Appendix C.2.

As a result of the difficulties faced, the ethics department granted permission for me to use existing datasets from external sources for my thesis. Following this approval, I identified a company that provided access to a dataset containing job advertisements in Country A. The company was discovered on Kaggle, an online platform that enables users to upload and share resources such as datasets and code notebooks (Kaggle, n.d.).

#### Dataset Information

The dataset was supplied in comma-separated value (CSV) format. It was downloaded and saved on a local password-protected computer and the OneDrive service that is accessible only to my school account, similar to the datasets in Chapter 4.

The dataset comprises 101,000 unique records gathered over a six-month period from March 2023 to August 2023. This dataset was sourced from 37 job boards in Country A, with only the data from Website A included in the dataset. Of these total records, Website A provided 22,100 records. This dataset contains only textual content in English without any images or emojis. It comprises 45 columns that include essential information such as job description, requirements, city, title, salary, and years of experience.

### 6.2.2 Data Processing

The dataset went through various preprocessing stages to make it suitable for annotation. The first step was to remove columns that contained metadata, duplicate information, personal details, and dataset status. Consequently, the number of columns was reduced from 45 to 11. The remaining columns comprised job title, city, job description, company name, company type, salary unit, salary range (both starting and ending), years of experience (both minimum and maximum) and industry. These columns were retained because they contain information that Organisation B uses to identify and investigate fraudulent jobs for human trafficking.

The next stage involved reducing the number of advertisements to 1,000 for annotation purposes. The decision to select 1,000 data points was made because of financial constraints and limited time. It was agreed with Organisation B that two experienced individuals from the organisation would assist with labelling the data and be compensated for their time, with efforts made to ensure minimal disruption to their work.

A series of queries were devised to select 1000 job advertisements for labelling out of the 22,100 job advertisements available for Website A. These queries aimed to guarantee that the chosen job advertisements contained the necessary information for experts to label the data accurately and to increase the likelihood of detecting advertisements related to human trafficking. The queries involved selecting job advertisements based on:

1. Jobs that are at high risk of human trafficking, such as domestic workers. Several keywords were used to identify these jobs such as "housekeeping", "cleaner", "nanny", and "domestic worker".

2. Jobs with language associated with fraudulent job advertisements for human trafficking. This shortlisted job descriptions includes text information such as "no experience required" and "youth".

3. Jobs with no experience required. The dataset had a column that provided details regarding the number of years of experience required, and this was used to select these types of jobs.

4. Job advertisements containing companies and websites that are known by Organisation B to advertise fraudulent job advertisements for human trafficking.

5. Jobs with salary information because this is one of the most important indicators used by experts. Not every job advertisement had the salary advertised.

These queries resulted in a CSV file with 1000 rows and 11 columns.

### 6.2.3 Data Annotation and Coding Scheme

The annotation process involved several iterations to label the dataset. Initially, a meeting was held with the director to determine the most suitable people to label the data. The Director decided that the project manager and data analyst would be the best candidates for the task, considering their relevant work experience and time availability. As mentioned, the project manager had seven years of experience, while the data analyst, who had previously worked as an investigator, had two years of experience. The following sections discuss the annotation guide that was employed to guide the labelling process, as well as the actual process of labelling itself.

### Annotation Guide

Before the annotation process was initiated, a meeting was held with the two experts to explain the annotation guidelines and to answer any questions they may have. The meeting also sought feedback on the annotation guidelines.

Annotators were required to label each job advertisement with the following labels:

1. Not Suspicious - This label was reserved for jobs that seemed legitimate and posed no reason for concern to annotators. Experts would select this option if they were confident that the job advertisement was not related to human trafficking.

2. Suspicious - This label applied to job advertisements that appeared to be illegitimate but there was insufficient evidence of human trafficking. Experts would select this option if they were not very confident that the job advertisement could be linked to human trafficking or be considered not suspicious.

3. High Risk of Human Trafficking - This label was applied to job advertisements that contained strong signs of human trafficking. Experts would select this option if they were confident that the job advertisement was linked to human trafficking.

For experts to label jobs as high-risk of trafficking or suspicious, they were asked to consider the indicators described in Chapter 5. A summary of the indicators was provided to the experts in the annotation guide, which is attached in Appendix C.1. Therefore, experts used these indicators during the annotation process. As mentioned in Section 5.3, some experts have also devised customised methods to identify instances of human trafficking beyond Organisation B's official set of indicators. Therefore, experts were allowed to use indicators not listed in the guideline, as well as any investigative techniques used in their daily work, to examine job advertisements.

| Kappa Score | Agreement |
|---|---|
| <0 | Less than chance agreement |
| 0.01 - 0.20 | Slight agreement |
| 0.21 - 0.40 | Fair agreement |
| 0.41 - 0.60 | Moderate agreement |
| 0.61-0.80 | Substantial agreement |
| 0.81 - 0.99 | Almost perfect agreement |

**Table 6.1:** Kappa Scores (Fendler et al., 2017)

### Iteration 1

In the first iteration, experts were given a month to label the 1000 data points. On average, each expert took two weeks to complete the task. Upon completion of the labelling process, it was discovered that 75 of the jobs were from the UK and the US and not from Country A. Further investigation revealed that these job advertisements shared similar city or town names with those in the US and the UK, leading to their misclassification as belonging to Country A. Five cities were identified as the source of this error. To rectify this issue, 75 new job advertisements were selected and added to the dataset, using the queries mentioned above.

Upon receiving the labelled job advertisements from both experts, the inter-rate reliability (IRR) was calculated to determine the level of agreement between them using Cohen Kappa (Bruijn, 2020; Widmann, 2020). Cohen Kappa is suitable for calculating the level of agreement with categorical data. The calculation showed a fair level of agreement, with a score of 0.27 based on the score table below. I opted to have another round of labelling to improve the IRR score to substantial agreement or almost perfect agreement. A high IRR score helps ensure that the data are accurate, consistent, and reliable (Mucinno, 2019; Parti, 2024).

### Iteration 2

A meeting with both experts was scheduled before starting the second labelling iteration. This meeting was scheduled to discuss the first labelling process and challenges, if any. During this meeting, both experts explained the data labelling process they used. It was observed that the project manager considered various

aspects of the job advertisement when reviewing the jobs. In contrast, the data analyst did not consider all aspects of the job and acknowledged that they were not as thorough as the project manager. The data analyst mainly focused on quantitative information, such as salary and number of years, and not qualitative aspects of the job advertisement, such as the job description. The project manager explained how he reviewed the advertisements, focussing on both the qualitative and quantitative attributes of the job. However, it was also noted that the project manager had considerable contextual knowledge about the locations and companies that helped them determine whether the advertisements were related to human trafficking.

During this iteration, additional steps were taken to ensure that all labelled data were from Country A. The IRR in the second iteration demonstrated a score of 0.03, indicating a slight agreement. The score decreased significantly from the first interaction. I observed that experts were unable to distinguish between high-risk human trafficking and suspicious labels. The project manager labelled more advertisements as having a high risk of human trafficking, whereas the data analyst labelled many job advertisements as suspicious. The data analyst confirmed their hesitation to label a job advertisement as "high risk of human trafficking" because of the lack of detailed investigation. To address these inconsistencies, data workshops were organised with experts to label the data.

### Data Workshops

These data workshops aimed to find a way to collaboratively label the dataset and resolve discrepancies. A total of 15 hours were spent labelling the data in seven data workshops.

The first workshop was used to discuss job posts and clarify the difference between "suspicious" and "high risk of human trafficking" labels. It was confirmed that the data analyst was not clear on the difference and hesitated to label job advertisements as high-risk despite sufficient evidence.

All workshops had a designated time during which experts were allowed to examine job advertisements independently. This was done to provide experts with

the opportunity to reconsider their opinions and modify them if necessary. Following the end of this period, we conducted a collective review of each job advertisement that was independently reviewed with each expert stating their label. In the event that the labels were similar, the job advertisement and the label were recorded in a separate spreadsheet. In the case of different labels, each expert would offer an explanation on the reasoning for selecting their label, and we would engage in a discussion to determine the appropriate label for the job advertisement. I facilitated discussions in which both experts were able to provide insight into the final label. The discussion process always began with the data analyst to ensure that they were heard and did not change their initial views on the job advertisement before the project manager communicated their views.

For the first two workshops, regardless of similarities or differences, each expert explained the reasoning behind labelling the data. This was beneficial because the project manager who was more experienced in this field provided valuable insight. The final five sessions ran more quickly as the level of agreement increased significantly, particularly in cases categorised as "not suspicious" and "high risk of human trafficking".

Upon reflection, the project manager recognised that they overlooked crucial information during the initial labelling process, while the data analyst did not thoroughly examine job advertisements associated with websites known to engage in human trafficking. Some job advertisements on Website A were posted by companies and other websites. Some of these websites were typically associated with human trafficking, and the data analyst tended to label these jobs as "suspicious", which was not always correct. At the end of the workshops, we produced a labelled dataset with 1000 data points which will be called "1K dataset" for the rest of the thesis.

The data workshops improved the quality of the data in several key ways. Firstly, they enhanced accuracy through collaboration by considering each expert's perspective, ensuring a more balanced and thorough labelling process. Secondly, the workshops allowed the experts to refine their understanding of key labels, particularly the distinction between "suspicious" and "high risk of human trafficking".

By resolving discrepancies in the labelling process, the overall quality of the dataset was significantly improved.

### 6.2.4 1K Dataset Statistics

The 1K dataset has 20244 distinct unigrams, 103019 distinct bigrams, and 164618 distinct trigrams. The total number of distinct characters was 125 characters. Unigrams are individual words or tokens, bigrams are pairs of consecutive words or tokens, and trigrams are sequences of three consecutive words or tokens.

Table 6.2 summarises the statistics for the labels.

| Label | Total | Average Word Length |
|---|---|---|
| Not Suspicious | 680 | 1781 |
| High Risk of Human Trafficking | 188 | 1560 |
| Suspicious | 132 | 1606 |

**Table 6.2:** Label Statistics

A visual representation of the top 20 unigrams of high-risk and suspicious advertisements is shown in Figure 6.1. Words such as "experience", "ensure", "work", "management", and "skills" are commonly found in these job advertisements.

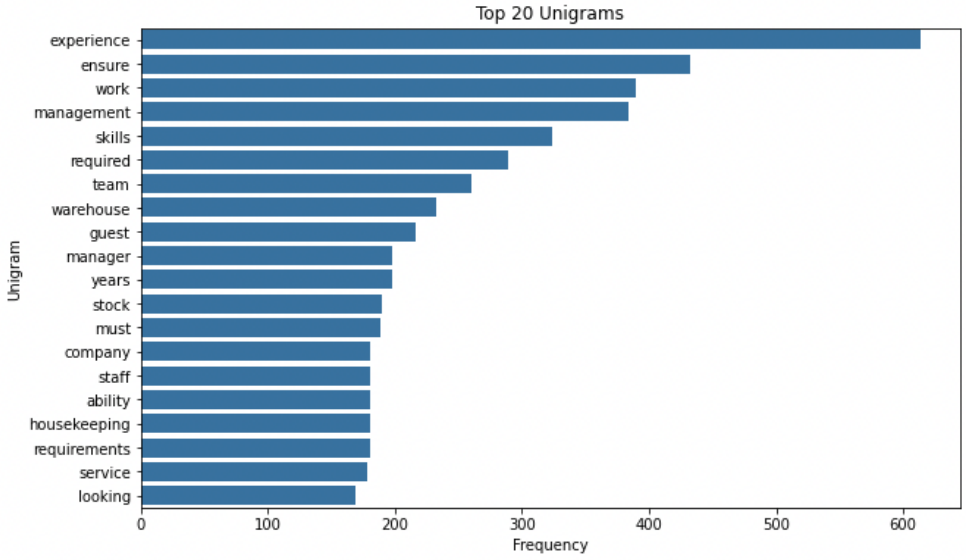


**Figure 6.1:** Top 20 Unigrams from Job Advertisements Labelled as "High Risk of Human Trafficking" and "Suspicious" (1K Dataset)

## 6.3 Classification: Benchmarking Various Approaches

This section focusses on classifying and evaluating the labelled dataset using various methods, including traditional machine learning models, pre-trained language models, and large language models. These approaches have proven effective in various disciplines, with certain methods demonstrating promising outcomes through recent advancements in the field of LLM, while others continue to showcase their effectiveness with pre-trained language models and traditional machine learning methods.

### 6.3.1 Traditional Machine Learning Approach

Classifying the dataset using traditional machine learning involved three primary stages: preprocessing the annotated data, vectorising the data using word embeddings, and employing machine learning models for classification. The main steps involved in classifying job advertisements using traditional machine learning approaches are summarised in Figure 6.2

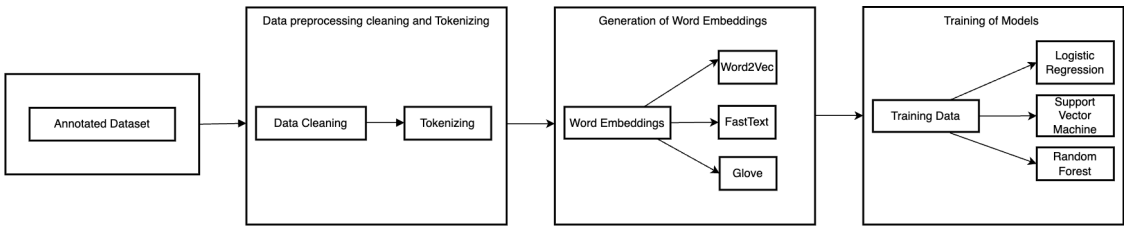


**Figure 6.2:** Overview of Traditional Machine Learning Approach

#### Pre-processing

The 1K dataset was pre-processed to ensure that it was in a suitable format. This involved removing columns containing personal information and integrating the remaining columns into the job description column, creating a single column that contained all relevant information about the job advertisement. The following step involved tokenising the job advertisement column using word tokenisation. Tokenisation in natural language processing involves breaking down text into units that can be easily understood by machine learning algorithms (Burchfiel, 2022). It

helps in understanding the context of the text and improves the accuracy of the other stages (Ashraf, 2024). Fraudulent job advertisements for human trafficking require an NLP approach that considers context because the advertisements are written with the aim of coercing victims, making tokenization an important step in this process. An example of how tokenization works for this task is as follows: each job advertisement is treated as an input that produces a sequence of tokens (words). Given a job advertisement $T$:

$$T = \text{"Remote work opportunity with high salary and flexible hours"}$$

Tokenization:

$$T \rightarrow \left\{ \begin{array}{l} \text{"Remote", "work", "opportunity",} \\ \text{"with", "high", "salary", "and"} \\ \text{"flexible", "hours"} \end{array} \right\}$$

This section used the Natural Language Toolkit (NLTK) word tokenizer, "word_tokenize" (NLTK, n.d.),.

**Vectorisation using Word Embeddings**

Word embeddings were chosen because they consider the context of information when vectorising data, unlike approaches such as Bag of Words and TF-IDF, which do not account for context (Tum, 2020). Word embedding models vectorise each word by assigning similar words similar representations, which are then learnt using a neural network (Brownlee, 2019). This section focusses on the use of the Word2Vec, FastText, and GloVe models.

**Word2Vec** Word2Vec was used for this approach due to its ability to produce high-quality embeddings that capture semantics when generating embeddings (Brownlee, 2019; Hoque, 2023). Word2Vec uses either the Continuous Bag-of-Words model (CBOW) or the Skip-gram model (Suri, 2022).

Given a target word $w_t$ and a context window of size $C$, the context words are $(w_{t-C}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+C})$ (Mikolov et al., 2013; Rong, 2014). CBOW predicts a word using the surrounding words; it takes an input X and aims to predict the output y. It predicts y according to the context.

**CBOW:** The objective is to maximize the average log probability (Mikolov et al., 2013; Rong, 2014):

$$\frac{1}{T}\sum_{t=C}^{T-C}\log P(w_t \mid w_{t-C},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+C}) \tag{6.1}$$

In contrast, with the Skip-gram model, given a target T, it predicts the surrounding words using the target word. The model is fed (X,y), with X representing the input while y representing the output.

**Skip-gram:** The objective to maximize is (Mikolov et al., 2013; Rong, 2014):

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-C\leq j\leq C\\ j\neq 0}}\log P(w_{t+j} \mid w_t) \tag{6.2}$$

Both **CBOW** and **Skip-gram** use softmax to convert the raw output into probabilities in this section (Mikolov et al., 2013; Rong, 2014).

$$P(w_O \mid w_I) = \frac{\exp(v'_{w_O}\cdot v_{w_I})}{\sum_{w=1}^{W}\exp(v'_w\cdot v_{w_I})} \tag{6.3}$$

Where $v_{w_I}$ and $v'_{w_O}$ are the input and output vector representations of $w$.

**FastText** FastText was chosen because of its ability to operate at the character level, which allows it to perform well with new or rare words, which is important in human trafficking, as traffickers always produce new words to prevent detection (Simonson, 2021; Subedi, 2018; Ganesan, n.d.). Similarly to Word2Vec, FastText uses CBOW or Skip-gram (Subedi, 2018; Ganesan, n.d.). It extends Word2Vec by representing the words as n-gram vectors.

For a word $w$ composed of character n-grams $\{g_1, g_2, \ldots, g_m\}$:

$$v_w = \sum_{i=1}^{m} v_{g_i}$$

The objective functions of CBOW or Skip-gram, described above, are then applied. Probabilities are calculated using the sum of n-grams (Mikolov et al.,

2013; Rong, 2014). Each n-gram $g_i$ has an associated vector $v_{g_i}$. The word vector $v_w$ is the sum of its n-gram vectors:

$$v_w = \sum_{i=1}^{m} v_{g_i}$$

**GloVe** GloVe was chosen because of its ability to capture the wider context of words, which can capture subtle meanings, making it suitable for the field of human trafficking (Pennington et al., 2014). GloVe generates the embeddings by generating word co-occurrence statistics from a text corpus.

Given a word-word co-occurrence matrix $X$, where $X_{ij}$ represents the frequency of word $j$ appearing in the context of word $i$, GloVe minimises the following objective function (Pennington et al., 2014):

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left( w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

Where:

- $w_i$ and $\tilde{w}_j$ are the word vectors.
- $b_i$ and $\tilde{b}_j$ are the biases.
- $f(X_{ij})$ is a weighting function that assigns less importance to infrequent co-occurrences.

Once word embeddings are generated after each approach above, each word $w_i$ in the job advertisement $T$ is represented as a vector $v_{w_i} \in \mathbb{R}^d$, where $d$ is the dimensionality of the embedding space. The dimensions used for Word2Vec and Fast2Text are $d = 100$ and for GloVe are $d = 300$.

### Combining Word Embeddings

This stage is important to represent the job advertisement as a single vector from the previous stages. The combined vector captures all the context and meanings of the entire advertisement. This helps machine learning algorithms to analyse and classify the text in the next stage.

To represent the entire job advertisement $T$ as a single vector $V_T$, the **averaging method** was used (V. Gupta et al., 2019):

$$V_T = \frac{1}{n} \sum_{i=1}^{n} v_{w_i}$$

### Machine Learning Classifier

Once the aggregated findings have been established $V_T$, a machine learning model is used to classify the jobs according to the three labels. Three machine learning algorithms were used to classify the job advertisements: logistic regression, support vector machine, and random forest.

Logistic regression was chosen due to its ability to provide clear and interpretable results, which is ideal to achieve explainability, an important factor in human trafficking cases, as previously explained (Kujlzb, 2024). The support vector machine and random forests were selected because they are effective with unbalanced data and help reduce overfitting, making them well suited for this type of problem where the data are not balanced (Sruthi, 2024; Breiman, 2001; Tabsharani, 2023).

**Logistic Regression** Given a set of job descriptions $K$, $K = \{(x_i, y_i)\}_{i=1}^{n}$, consisting of $n$ training samples, $x_i$ as input features and $y_i$ as output features with three labels, $y_i \in \{1, 2, 3\}$, multi-class classification is achieved using the One-vs-One (OvO) method (Brownlee, 2021). OvO was chosen because of its ability to handle imbalances with different classes (Wadhwatanya, 2021). This method trains a separate classifier for each pair of classes.

For each pair class $(j, k)$, we define $y_{ijk}$ as:

$$y_{ijk} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{if } y_i = k \end{cases}$$

The objective function for each classifier $(j, k)$ is as follows (Dobson and Barnett, 2018; Saini, 2024):

$$\min_{\beta_{jk}} \sum_{i \in \{y_i = j \text{ or } y_i = k\}} \Big[ y_{ijk} \log h_{\beta_{jk}}(x_i) + (1 - y_{ijk}) \log(1 - h_{\beta_{jk}}(x_i)) \Big]$$

Where:

$$h_{\beta_{jk}}(x_i) = \frac{1}{1 + \exp(-\beta_{jk}^T x_i)}$$

Here, $y_{ijk}$ is a binary indicator (1 if class $j$, 0 if class $k$) for the pair of classes being compared.

In this thesis, we also used class weight (Scikit Learn, n.d.) to help resolve class imbalance issues with the dataset and prevent the model from being biased toward the "non-suspicious" label, which occurs more frequently. This would be a similar setup in a real-life setting where 'non-suspicious' job advertisements are more prevalent on job boards than job advertisements related to human trafficking (Rivas et al., 2024). Furthermore, L2 regularisation is used to prevent overfitting with the model and improve generalisation of new data (M, 2020; RITHP, 2023).

**Support Vector Machine (SVM)** OvO is also used to handle multi-class classification with three labels. This method involves training a separate SVM classifier for each pair of classes (Brownlee, 2021). Given three labels, $y_i \in \{1, 2, 3\}$, the OvO approach solves an optimisation problem for each pair.

For each pair, $(i, j)$, optimisation problem can be defined as:

$$\min_{w_{ij}, b_{ij}, \xi_{ij}} \quad \frac{1}{2} \|w_{ij}\|^2 + C \sum_{m=1}^{n_{ij}} \xi_{ij,m} \tag{6.4}$$

(Piccialli and Sciandrone, 2022; Chauhan et al., 2019)

Subject to:

$$y_{ij,m}(w_{ij} \cdot x_{ij,m} + b_{ij}) \geq 1 - \xi_{ij,m} \tag{6.5}$$

$$\xi_{ij,m} \geq 0 \tag{6.6}$$

Where:

- $w_{ij}$: Weight vector for classifier distinguishing class $i$ and class $j$
- $b_{ij}$: Bias term for classifier $(i, j)$
- $\xi_{ij,m}$: Slack variables for classifier $(i, j)$
- $y_{ij,m} \in \{-1, 1\}$: Labels for the pairwise classifier

Each classifier $C_{ij}$ produces a decision function $f_{ij}(x)$ which determines whether a given sample $x$ belongs to class $i$ or class $j$ (Piccialli and Sciandrone, 2022; Chauhan et al., 2019). The results of these pairwise classifiers are combined through a voting mechanism to produce a final multi-class prediction. The class balance feature and L2 regularisation are also used for SVM to help prevent bias and overfitting.

**Random Forest** This algorithm builds a forest using multiple decision trees that operate as an ensemble (DeAngelo et al., 2019; Yiu, 2019). Decision trees $T$ divide the data into different groups using if and else statements based on the features, which eventually lead to a predicted class (Koehrsen, 2018). Each split was used to separate the data into homogeneous groups.

For each tree $t$, the objective is to minimise the Gini impurity at each split (Breiman, 2001; Saini, 2024):

$$G(t) = 1 - \sum_{i=1}^{c} p_i^2 \tag{6.7}$$

where $p_i$ is the proportion of samples belonging to class $i$ in the node.

The RF builds thousands of decision trees, denoted as $T_1, T_2, \ldots, T_B$ where $B$ is the total number of trees (Breiman, 2001; Saini, 2024). Each tree is constructed by using a random subset of data (with a replacement called bootsraping) and features.

The final prediction for a job advertisement is obtained by averaging the predictions of all trees by taking a majority vote (Breiman, 2001; Saini, 2024).

$$\hat{y} = \text{mode}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M) \tag{6.8}$$

where $\hat{y}_i$ is the prediction of the $i$-th tree.

In this thesis, the RF model incorporated the class weight parameter to address class imbalance and mitigate bias. Additionally, two hundred trees were used, as they provided the optimal performance based on multiple experiments. Furthermore, the random state variable was set to 42 to ensure the reproducibility of all trees.

**Experiment Results**

The performance of the models was assessed through cross-validation, which involves dividing data into training and testing sets multiple times. Cross-validation is known to reduce bias, which is a common issue in anti-trafficking software (Brownlee, 2023; Lyashenko and Jha, 2024). In particular, Stratified K-Fold Cross-Validation was employed, dividing the data into five equal parts. This method was chosen because it ensures that each part contains a similar proportion of job advertisements for each class, which improves the model's ability to generalise and increases its reliability (Nagaraj, 2023). The average performance across all five folds was calculated to determine the overall performance of the model.

Several performance metrics were used to evaluate the models, including accuracy, precision, recall, and the F1 score (Bajaj, 2023; Jain, 2024). Accuracy quantifies the ability of the model to predict all instances in the dataset. Precision evaluates the model's ability to accurately predict positive instances among all the identified positive instances. In contrast, recall measures the proportion of actual positive instances that are correctly identified. The F1 score, which is the harmonic mean of precision and recall, provides a balanced assessment of both metrics.

Tables 6.3 and 6.4 summarise the performance of the algorithms. The vector and window sizes for Word2Vec and FastText were determined based on the results of multiple experiments.

| Iterations | Model | Vector Size | Window Size | Embedding Algorithm |
|---|---|---|---|---|
| 1 | Word2Vec | 100 | 5 | Skip-gram |
| 2 | Word2Vec | 100 | 5 | CBow |
| 3 | FastText | 100 | 5 | Skip-gram |
| 4 | FastText | 100 | 5 | CBow |
| 5 | Glove | 300 | 5 | Common Crawl |

**Table 6.3:** Word Embeddings

| Experiment | Algorithm | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| 1 | Logistic | 0.788 | 0.744 | 0.788 | 0.768 |
| 1 | SVM | 0.734 | 0.649 | 0.734 | 0.654 |
| 1 | RF | 0.720 | 0.666 | 0.720 | 0.663 |
| 2 | Logistic | 0.784 | 0.751 | 0.784 | 0.737 |
| 2 | SVM | 0.727 | 0.646 | 0.727 | 0.639 |
| 2 | RF | 0.716 | 0.662 | 0.716 | 0.656 |
| 3 | Logistic | 0.785 | 0.771 | 0.785 | 0.763 |
| 3 | SVM | 0.736 | 0.652 | 0.736 | 0.655 |
| 3 | RF | 0.708 | 0.659 | 0.708 | 0.652 |
| 4 | Logistic | 0.739 | 0.661 | 0.739 | 0.678 |
| 4 | SVM | 0.707 | 0.579 | 0.707 | 0.605 |
| 4 | RF | 0.692 | 0.630 | 0.692 | 0.637 |
| 5 | Logistic | 0.730 | 0.730 | 0.730 | 0.725 |
| 5 | SVM | 0.732 | 0.643 | 0.732 | 0.653 |
| 5 | RF | 0.717 | 0.667 | 0.717 | 0.652 |

**Table 6.4:** Performance Comparison of Different Algorithms across Multiple Experiments.

### 6.3.2 Pre-trained Neural Language Models

This section discusses the process of classifying job advertisements using three pre-trained neural language models: BERT, ALBERT, and RoBERTA. The primary steps involved in these methods include preprocessing the annotated information, creating word embeddings, and training the data.

The BERT model was selected for its superior performance compared to Word2Vec in NLP tasks (L. Gupta, 2020). This is largely because BERT understands the context of each word in relation to its surrounding words (context-dependent) and comprehends language semantics through extensive training on a vast corpus of text. Furthermore, BERT can effectively manage out-of-vocabulary words (L. Gupta, 2020; Bhakta, 2024). Human traffickers are known to use words that are not in

standard vocabularies to communicate information, making the use of BERT ideal. The ALBERT model, a simplified version of BERT, was chosen for its efficiency and its ability to reduce memory usage and increase training speed, which makes it ideal when computational resources are limited (Joy et al., 2022; Lan et al., 2019). RoBERTa, built on the BERT architecture, was selected because it was trained on larger and more diverse datasets than BERT, enabling it to capture more complex patterns in the data (Kacprzak, 2024; Jorge, 2023). These models are able to handle complex tasks in natural language, which will be crucial in identifying subtle indicators of human trafficking (Islam et al., 2024; Vaswani et al., 2017; Devlin et al., 2018). The main steps involved in PLMs are illustrated in Figure 6.3.

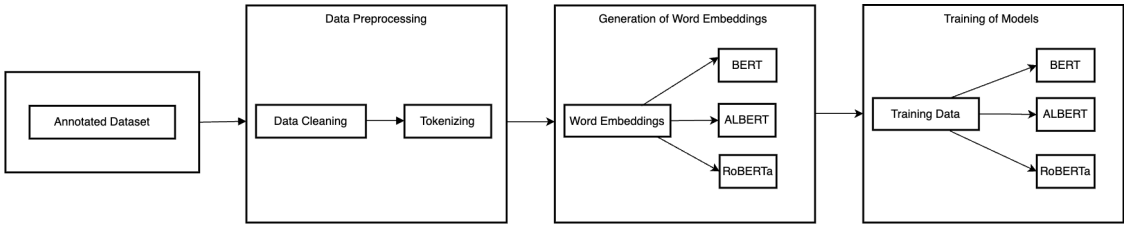


**Figure 6.3:** Overview of Pre-trained Neural Language Models Approach

### Preprocessing

This section follows preprocessing steps similar to the traditional machine-learning approaches described above. However, for PLMs, each model employs a custom tokenizer to convert the data into a form that the models can understand. The tokenizers used were BertTokenizer, RobertaTokenizer, and AlbertTokenizer.

The BertTokenizer and AlbertTokenizer use WordPiece, a subword tokenizer that breaks words into smaller units, allowing the tokenizer to handle words that are not in the vocabulary, which improves the performance of NLP tasks (V. Khanna, 2024). RoBERTa uses Byte-Pair Encoding (BPE), a compression algorithm that merges the most common character pairs in a vocabulary to represent text more efficiently (C. Khanna, 2021; Sanbit, 2024). This approach helps balance the size of the vocabulary while retaining valuable information. It is particularly beneficial

for OOV words, as BPE breaks down words into smaller units, allowing the model to process and understand them effectively (Sanbit, 2024).

For example, in WordPiece, given a job advertisement $T$:

$$T = \text{"Remote work opportunity with high salary and flexible hours"}$$

Tokenization:

$$T \rightarrow \left\{ \begin{array}{l} \text{"Remote", "work", "opportunity", "with",} \\ \text{"salary", "and", "flex", "\#\#ible", ".",} \end{array} \right\}$$

For BPE, given the job advertisement $T$. Tokenization:

$$T \rightarrow \left\{ \begin{array}{l} \text{"Re", "mo", "te", "work", "op", "port"} \\ \text{"unity", "with", "high", "salary", "and"} \\ \text{"flex", "ible", "hours", "."} \end{array} \right\}$$

**Generating Word Embeddings**

The PLMs used in this section have their own built-in embedding capabilities that are the core components of their architectures.

**BERT Embedding Layers** BERT generates word embeddings from the input text through three main steps: token embeddings, position embeddings, and segment embeddings. With token embedding, the text is mapped to dense vector representations.

- Let $V$ be the size of the vocabulary.
- Let $d$ be the dimensionality of the embeddings.
- The token embedding matrix $E_{\text{token}}$ is of size $V \times d$.

Given an input x, let $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ where $x_i$ is the token at position $i$. The token embedding for $e_{\text{token},i}$ for the token $x_i$ is given by (Adith, n.d.; Raman, 2023):

$$e_{\text{token},i} = E_{\text{token}}[x_i]$$

For a token $x_i$ in BERT, the embedding is:

$$e_{\text{token},i} = E_{\text{token}}[x_i]$$

where $E_{\text{token}} \in \mathbb{R}^{V \times d}$.

Position embeddings help the model understand the order of tokens in a sequence (Adith, n.d.; Raman, 2023). The position embedding $e_{\text{pos},i}$ for the token at position $i$ is given by:

$$e_{\text{pos},i} = E_{\text{pos}}[i]$$

Segment embeddings are used when the input consists of a pair of sentences (Adith, n.d.; Raman, 2023). The model differentiates between these two sequences. For an input sequence with corresponding segment IDs $\mathbf{t} = [t_1, t_2, \ldots, t_n]$ where $t_i$ is the segment ID (0 or 1) at position $i$:

The segment embedding $e_{\text{seg},i}$ for the token at position $i$ is given by (Adith, n.d.; Raman, 2023):

$$e_{\text{seg},i} = E_{\text{seg}}[t_i]$$

For each input token, the final representation is the sum of token embeddings, position embeddings, and segment embeddings. The combined embedding $e_i$ is given by:

$$e_i = e_{\text{token},i} + e_{\text{pos},i} + e_{\text{seg},i}$$

Token embeddings are beneficial for understanding words in the context in which they are used, helping to identify subtle cues in job advertisements that may indicate a risk of human trafficking. Positional embeddings, on the other hand, help interpret the meaning of a sentence based on word order, thus improving the understanding of the overall meaning of a job advertisement (Adith, n.d.; Raman, 2023). Segment embeddings help to categorise different parts of the text, such as job descriptions and requirements, which can assist in detecting unusual patterns (Adith, n.d.; Raman, 2023).

**ALBERT Embedding Layers** ALBERT has a similar process to BERT but does improve efficiency by introducing optimisation techniques to the model. The main difference is with token embedding; in ALBERT, the model reduces the size of a vocabulary embedding matrix using a factorised embedding parameterisation.

- Let $V$ be the size of the vocabulary.
- Let $d$ be the dimensionality of the hidden layers.
- Let $d_e$ be the embedding size, where $d_e < d$.
- The token embedding matrix $E_{\text{token}}$ is of size $V \times d_e$.
- A linear projection $W_{\text{proj}}$ is applied to project embeddings from size $d_e$ to size $d$.

For an input token $x_i$, the token embedding is (Lan et al., 2019):

$$e_{\text{token},i} = W_{\text{proj}} \cdot E_{\text{token}}[x_i]$$

where $E_{\text{token}} \in \mathbb{R}^{V \times d_e}$ and $W_{\text{proj}} \in \mathbb{R}^{d_e \times d}$.

ALBERT calculates the position and embedding segment in an approach similar to the BERT model above. The final representation for each input sequence is also calculated in a manner similar to BERT.

**RoBERTa Embedding Layers** RoBERTa embedding architecture is similar to BERT. However, it differs in that it does not use segment embeddings as part of its pre-training task (Y. Liu et al., 2019; Huang et al., 2020). However, the architecture still supports segment embeddings if necessary. The token embedding, the position embedding, and the combined embedding can be represented in a similar manner.

### Pre-trained Language Models

This section discusses the training process of the models after generating the embeddings.

**BERT** BERT uses a transformer architecture consisting of multiple self-attention and feed-forward neural networks. For each transformer layer.

For each transformer layer $l$ in the $L$ layers:

**Multi-Head Self-Attention:**

For each attention head $h$ (Devlin et al., 2018):

$$Q_h = EW_h^Q, \quad K_h = EW_h^K, \quad V_h = EW_h^V$$

$$A_h = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right)$$

$$H_h = A_h V_h$$

**Concatenate and project (Devlin et al., 2018):**

$$H = [H_1, H_2, \ldots, H_H]W^O$$

**Feed-Forward Network (Devlin et al., 2018):**

$$\text{FFN}(H) = \text{ReLU}(HW_1 + b_1)W_2 + b_2$$

**Layer Normalization and Residual Connections (Devlin et al., 2018):**

$$O_1 = \text{LayerNorm}(E + H)$$

$$O_2 = \text{LayerNorm}(O_1 + \text{FFN}(O_1))$$

**Output Layer:**

Final hidden state of [CLS] token (Devlin et al., 2018):

$$h_{\text{CLS}} = O_2[L]$$

**Training Objectives:**

Cross-Entropy Loss for Classification: For each example $m$:

$$\hat{y}_m = \text{softmax}(W_{\text{cls}} h_{\text{CLS},m} + b_{\text{cls}})$$

Loss:

$$L_{\text{cls}} = -\sum_{m} \sum_{c=1}^{C} y_{m,c} \log(\hat{y}_{m,c})$$

**ALBERT** ALBERT uses a similar training approach to BERT, but uses a technique of passing parameters across the layers to help reduce the size of the model (Lan et al., 2019). Therefore, it has significantly fewer parameters than BERT, but this does not significantly affect performance.

**RoBERTa** RoBERTa uses the same architecture as BERT, incorporating multi-head self-attention and feed-forward neural networks, but it also introduces dynamic masking during pre-training, which involves changing the masked token at each training iteration (Efimov, 2023). Dynamic masking is beneficial because it prevents overfitting to a particular masking pattern, encourages the learning of more generalized text representations, and allows access to more data by masking different tokens (Y. Liu et al., 2019). This increased diversity during training enhances the model's overall performance (Y. Liu et al., 2019; Efimov, 2023).

### Experiment Results

These three models were also evaluated using stratified cross-validation to improve reliability.

| Model | Version | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| BERT | bert-base-uncased | 0.8290 | 0.8110 | 0.8290 | 0.8174 |
| ALBERT | albert-base-v2 | 0.8730 | 0.8727 | 0.8730 | 0.869 |
| RoBERTa | roberta-base | 0.8790 | 0.8841 | 0.8790 | 0.875 |

**Table 6.5:** Performance Comparison of BERT, AlBERT, and RoBERTa on the 1K Dataset.

### 6.3.3 Large Language Models Approach

This section focusses on the application of two widely used large language models, GPT-3.5 and GPT-4. These models are capable of producing better results than pre-trained language models due to their extensive training in vast amounts of data and their advanced ability to understand and generate natural language (Minaee et al., 2023; Kalyan, 2023). Although OpenAI, the creators of these models, do not provide API access to fine-tune most GPT models, they still demonstrate impressive performance in zero-shot learning (Minaee et al., 2023; Kalyan, 2023). The main steps involved in this process are illustrated in Figure 6.4.

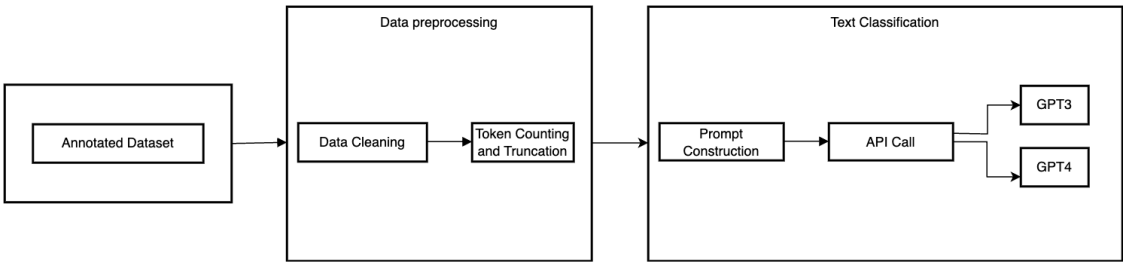


**Figure 6.4:** Overview of Large Language Models Approach

#### Preprocessing

The preprocessing step primarily entailed cleaning the data in a manner similar to the other two approaches, removing unnecessary columns and merging columns into a single column, job description.

However, for LLMs unlike the other approaches, token counting and truncation are performed to ensure that the prompt used for classification fits within the token limits of the models. This process helps prevent errors by truncating any text that exceeds the token limit.

#### Large Language Model Classification

This section discusses the prompts used for classification and the process of training the models.

**Prompt Construction** During this stage, a prompt was designed to classify job advertisement texts into three categories: "not suspicious", "suspicious", and "high risk of human trafficking".

The prompt structure included both a prefix and a suffix:

- The prefix provided task-specific instructions and outlined the classification labels.
- The suffix specified where the model's response should begin to ensure structured outputs.

Due to the high costs associated with running multiple experiments, only a simple exploratory analysis of the prompts was conducted.

**GPT-3** GPT-3 is widely recognised as one of the first models to demonstrate strong performance in various tasks of natural language, such as generation, summarisation, and classification (Kublik and Saboo, 2022). It has 175B parameters and is trained from various data sources, such as books and websites (Kalyan, 2023). The introduction of GPT-3.5 addressed some of the limitations of its predecessor, such as the generation of harmful content and the failure to follow user instructions (Kalyan, 2023). This model was selected because of its ability to handle unseen tasks with high proficiency (Kalyan, 2023).

GPT-3.5 uses Byte-Pair Encoding (BPE) for tokenization and implicitly generates word embeddings. Both Pretrained Language Models (PLMs) and Large Language Models (LLMs) are based on transformers and share core architectural components. The key difference is that GPT models focus on generating the next token in a sequence based on the previous tokens, while BERT models aim to understand the context of tokens by simultaneously considering both the preceding and following token (Devlin et al., 2018; Kalyan, 2023).

**GPT-4** GPT-4 is an enhanced version of GPT-3 and GPT-3.5, capable of handling both text and images, and has been trained on 45 gigabytes of data, compared to 17 gigabytes for GPT-3.5, which is only capable of handling text (Kujlzb, 2024; Kalyan, 2023). Although GPT-4 shares a similar architecture with GPT-3.5, it has a larger number of parameters, making it more powerful.

GPT-4's capabilities include handling longer texts due to its larger context size compared to earlier GPT models, as well as offering more safety features and advanced functionality that are better aligned with user prompts (Kujlzb, 2024; Kelly, 2024). GPT-3.5 lacks recent data, limiting its ability to perform tasks that require information from 2021 (Kujlzb, 2024; Kelly, 2024). These enhancements enable GPT-4 to capture subtle cues and hidden intent in complex and elaborate job descriptions that GPT-3.5 might miss.

### Experiment Results

These three models were also evaluated using cross-validation specifically stratified cross-validation.

| Model | Algorithm | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| GPT-3 | gpt-3.5-turbo-instruct | 0.84 | 0.86 | 0.84 | 0.81 |
| GPT-3 | gpt-3.5-turbo-0125 | 0.78 | 0.82 | 0.78 | 0.74 |
| GPT-4 | gpt-4 | 0.73 | 0.72 | 0.73 | 0.65 |
| GPT-4 | gpt-4-turbo | 0.71 | 0.73 | 0.71 | 0.63 |

**Table 6.6:** Performance Comparison of GPT-3.5 and GPT-4 on the 1K Dataset.

## 6.4 Expert Feedback

The goal of this section is to classify unlabelled job advertisements on Website A using the best performing model and to validate its results by comparing the generated labels with those provided by experts. The aim was to achieve an almost perfect Inter-Rater Reliability (IRR) score between the two datasets. Achieving a nearly perfect score between the model and expert labelling was crucial to

validating the model's results and ensuring its reliability in labelling unlabelled data (Mucinno, 2019).

One hundred random data points were selected from the 22,100 job advertisements available on Website A. Steps were taken to ensure that these 100 data points were not included in the initial 1K training dataset to avoid overfitting and bias. The following sections discuss the process of labelling with the best-performing model, workshops with experts, and a discussion of the results.

**Labelling with Best Performing Model**

After evaluating traditional machine learning, PLM, and LLM in the 1K data set, it was determined that RoBERTa achieved the most favourable results among all models tested. With an accuracy rate and an F1 score of 0.88, RoBERTa outperformed. RoBERTa was deployed and used to label the 100 unlabelled job advertisements. These job advertisements were saved in CSV format.

**Labelling with Experts**

Two workshops were held, similar to those conducted during the labelling of the initial 1K dataset. In each workshop, experts independently labelled the dataset and engaged in detailed discussions of each row to evaluate the similarity and dissimilarity of the labels. If the labels for a row were found to be dissimilar, the row was reviewed again until a consensus label was reached.

The job advertisements classified by the model were also considered, particularly those identified as "high risk of human trafficking" or "suspicious". The experts were asked to reassess these job advertisements if there were discrepancies between the model's labels and the experts' labels. This process aimed to determine whether the model identified high-risk and suspicious jobs that experts might have missed.

These workshops helped improve the quality and accuracy of the data labels while maintaining time efficiency. In addition, the workshops provided valuable feedback from the experts on jobs labelled as high-risk or suspicious and, in some cases, impacted their decisions. This is discussed in Section 6.6.

**Results**

The level of agreement between the RoBERTAa-labelled dataset and the expert-labelled dataset was calculated using the IRR method, percentage agreement. This type of agreement is calculated by dividing the total number of agreements by the total number of observations (Bobbitt Zach, 2021).

Table 6.7 compares the labels assigned by experts and the RoBERTa model. It presents the number of job advertisements classified under each label by both the model and the experts, followed by the corresponding agreement score. The overall agreement score was 91%, which is considered excellent according to the percentage agreement method. This places the performance of the model in the highest category, indicating near-perfect alignment with expert classifications (Parti, 2024).

| Jobs | Experts | RoBERTa | Agreement Score |
|---|---|---|---|
| Not Suspicious | 93 | 94 | 95.74% |
| High Risk of Human Trafficking | 3 | 2 | 50% |
| Suspicious | 4 | 4 | 25% |

**Table 6.7:** Validating Roberta Results

## 6.5 Trafficking Recruitment 10K Dataset

After achieving appropriate agreement between the datasets, RoBERTa was used to 10000 unlabelled job advertisements, forming the 10K dataset. This dataset is useful for comparing the distribution and statistics of the data with the 1K dataset. Such a comparison can help determine whether similar insights, particularly regarding high-risk and suspicious data, can be observed across both datasets. The 10K dataset was also created to help researchers and other stakeholders conduct experiments that can aid in the development of solutions.

This dataset should not be considered as absolute truth, but rather as a starting point to create a more reliable dataset with the help of experts through validation techniques. It is important to note that the 10,000 labelled job advertisements did not contain any job advertisements in the 1K dataset. The 10K dataset has

52629 distinct unigrams, 609983 distinct bigrams, and 1581156 distinct trigrams. The statistics for the dataset are presented in Table 6.8.

| Label | Total | Average Word Length |
|---|---|---|
| Not Suspicious | 8793 | 2307 |
| High Risk of Human Trafficking | 325 | 2818 |
| Suspicious | 892 | 2298 |

**Table 6.8:** Label Statistics

A visual representation of the top 20 unigrams of job advertisements labelled as "high risk of human trafficking" and "suspicious" is shown in Figure 6.5. Five of the most common words in these job advertisements included "experience", "work", "team", "skills", and "management".

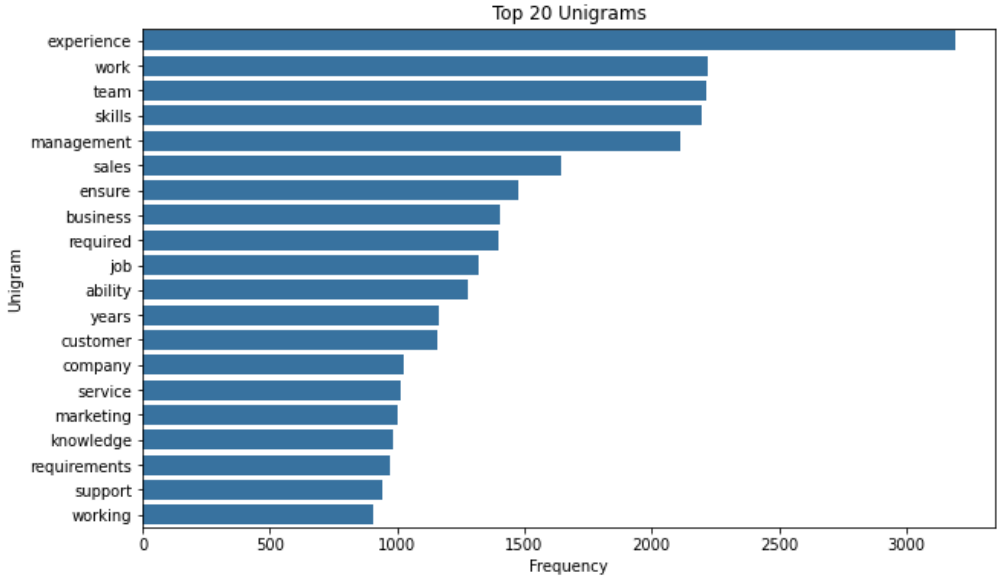


**Figure 6.5:** Top 20 Unigrams from Job Advertisements Labelled as 'High Risk of Human Trafficking' and 'Suspicious' (10K dataset)

## 6.6 Discussion

This section provides a summary of the findings and discusses the implications of these findings, as well as the results of the various approaches used to classify the data. It also includes an analysis of the job advertisements in the 1K dataset and examines the role of the P3D Framework in this work.

### 6.6.1 Summary of Findings

This chapter describes the process of creating an expert-labelled dataset, which was later used to train several models for benchmarking purposes. To the best of my knowledge, no prior work has been done on developing a dataset for fraudulent job advertisements in the context of human trafficking in Country A. Additionally, no existing baselines are currently available for this specific problem. Among the machine learning models trained, RoBERTa, a pre-trained neural language model, achieved the highest level of performance. Subsequently, RoBERTa was used to label 10,000 job advertisements for research purposes. These advertisements were labelled after the model validated the results using unlabelled job advertisements. The validation was carried out through expert workshops and by calculating the IRR between 100 RoBERTa-labelled job advertisements and the same 100 experts-labelled jobs. The IRR score was 0.91, indicating a nearly perfect agreement.

### 6.6.2 Analysis of Classification Results

AI methods have shown significant potential to address the challenges that arise with human trafficking recruitment on online platforms. This study evaluated the performance of various AI approaches for classifying online job advertisements for human trafficking.

#### Traditional Machine Learning Approaches

Traditional machine learning approaches initially struggled, particularly with the "suspicious" label, likely due to the complexity and ambiguous nature associated with it. However, performance improved significantly after incorporating word embeddings generated by Word2Vec, FastText, and GloVe, highlighting the importance of effective feature representation. Further improvement was observed when class imbalance was addressed. Tackling class imbalance is crucial, especially since certain classes are under-represented, such as in this case with the "suspicious" and "high risk of human trafficking" labels (Japkowicz and Stephen, 2002). The imbalance was addressed by adjusting the class weights.

Combining the "suspicious" and "high risk of human trafficking" labels can also help with class imbalance because it provides a larger sample size for the combined class and reduces the disparity with the "not suspicious" label. This approach would be appropriate because experts would be required to investigate job advertisements labelled as either "suspicious" or "high risk of human trafficking". However, it is advisable to separate these labels in situations where resources are limited to ensure that more attention and resources are directed toward investigating high-risk job advertisements, which are more critical.

**For future work using traditional machine learning approaches**, it is recommended to address the class imbalance to avoid bias in model predictions, mitigate poor performance associated with the suspicious and high-risk labels, and improve overall model accuracy (Japkowicz and Stephen, 2002; He and Garcia, 2009; Abhinav, 2023). Various techniques should be considered for balancing class weights, such as oversampling, undersampling, adjusting class weights, and cost-sensitive learning (Japkowicz and Stephen, 2002; He and Garcia, 2009; Abhinav, 2023). The appropriate method should be selected based on the specific classifier and the context of the real world application (Japkowicz and Stephen, 2002).

### Pre-trained Neural Language Models

Pre-trained language models tested in this thesis outperformed traditional methods, particularly in handling complex labels. Although BERT demonstrated average results on the "suspicious" label, both RoBERTa and ALBERT achieved moderately good results. BERT achieved an F1 score of 0.57 while RoBERTa and ALBERT achieved a score of 0.74 and 0.76 respectively. This suggests that providing more training examples, as discussed earlier, could improve performance.

The comparable performance of ALBERT and RoBERTa with challenging labels highlights their robustness in handling such tasks. Transformer-based models have consistently achieved state-of-the-art performance in complex text classification due to their ability to capture contextual information and maintain relationships between words and phrases effectively (Islam et al., 2024; Vaswani et al., 2017;

Devlin et al., 2018). Furthermore, neural networks are generally less affected by class imbalances, which may explain their strong performance in this context (Japkowicz and Stephen, 2002). Compared to other deep learning approaches in the human trafficking domain (Tong et al., 2017; Wang et al., 2020; Esfahani et al., 2020), RoBERTa achieves state-of-the-art results in terms of precision, recall and F1 score.

**To further improve the accuracy of pre-trained neural language models**, several strategies can be implemented. Hyperparameter tuning, particularly through techniques such as genetic algorithms, has been shown to enhance the performance of models like BERT and RoBERTa, especially for classification tasks (Sobhanam and Prakash, 2023). Furthermore, using larger variants of these models, which contain more layers and parameters, can also lead to improved performance (Vaswani et al., 2017; Kaplan et al., 2020). However, it is important to note that larger models require significantly more computational resources compared to their base versions (Kaplan et al., 2020).

### Large Language Models

Language models (LLMs), particularly GPT-3.5 turbo-instruct, demonstrated strong performance even in zero-shot settings, where no specific pre-training is conducted. This highlights their adaptability and potential for application in the field of human trafficking detection, aligning with previous studies that indicate that LLMs can achieve state-of-the-art results in zero-shot, few-shot, and one-shot settings (Chae and Davidson, 2023; Brownlee, 2019; Z. Wang et al., 2023).

GPT-3.5 outperformed GPT-4 in this context, which can be attributed to various factors, including differences in model architecture and dataset optimisation. GPT-3.5 turbo-instruct achieved the highest accuracy for the "suspicious" label, a challenging category for other approaches, with an impressive 91% accuracy. However, its performance was less robust for the "high risk of human trafficking" label, with a recall of 29% and an F1 score of 44%, despite a high precision score of 91%. Similar trends of poor recall and very high precision were observed in the other LLM variants.

Low performance on high-risk labels highlights the limitations of LLMs in tasks requiring deeper reasoning and understanding, such as this one (Brown et al., 2020; Sun et al., 2023; Alizadeh et al., 2023). Furthermore, LLMs tend to underperform compared to pre-trained neural language models due to the constraints on context learning imposed by token limitations (Sun et al., 2023). The low recall associated with the high-risk human trafficking label may also be due to the high decision threshold used by LLMs, leading them to make predictions only when they are highly confident.

**To enhance the overall performance of the LLMs**, it is recommended that the models be fine-tuned on the dataset (Brownlee, 2019; Alizadeh et al., 2023). In addition, experimenting with different prompts, particularly for the "High Risk of Human Trafficking" label, is advised to potentially surpass the results of models like RoBERTa.

While LLMs offer significant potential, they must be used cautiously due to the risk of data misuse by organizations and security breaches if protective measures are bypassed. Therefore, it is crucial to evaluate potential negative consequences before deployment to ensure responsible and secure usage.

### 6.6.3 Analysis of Job Advertisements

Analysis of job advertisements showed that advertisements labelled "high risk of human trafficking" tend to have shorter descriptions than those without this label. Generally, legitimate employers provide more comprehensive information about job openings in their advertisements, which explains why "not suspicious" job advertisements have the longest descriptions. Traffickers are less forthcoming with details because the jobs do not exist, which can result in shorter descriptions. Jobs categorised as "suspicious" fall between these two categories in terms of length. This is likely because experts are less certain about these positions, and the category serves as a neutral category between the other two labels.

Compared to the 10K dataset, the 'high risk of human trafficking' label is associated with the longest job descriptions, while the 'Not suspicious' label

corresponds to the shortest descriptions. This pattern differs from the 1K dataset, further emphasising the need to validate the dataset. Furthermore, when comparing the top 20 unigrams for high-risk and suspicious labels between the 1K and 10K datasets, it is observed that they share 9 common words. These common words, such as "experience", "skills", "team", and "ensure", are frequently used in job advertisements, which may explain their high occurrence. More research is needed to create a list of commonly used words by human traffickers in job advertisements.

In the validation set, the performance of the RoBERTa model was evaluated for its ability to classify job advertisements. Two job advertisements were initially assessed by experts as "not suspicious", but were classified as "High Risk of Human Trafficking" and "Suspicious" by the model. Upon further review, the experts agreed with the model's results, acknowledging that they had overlooked important information and that the classifications were accurate. High-risk and suspicious job advertisements posted on job boards are more difficult to evaluate than those posted on social media platforms. Traffickers attempt to comply with higher formality standards on job boards, which contrasts with the more casual nature of social media platforms. The use of automated methods can provide an additional layer of assistance to experts in identifying high-risk and suspicious cases that experts may overlook.

In certain cases, experts did not agree with the model outcomes, highlighting the importance of retaining experts in the decision-making process. The model did not detect certain job advertisements classified by experts as "high risk of human trafficking" and "suspicious" in the validation set. Furthermore, of the seven cases that experts classified as high-risk or suspicious, the model had a 43% accuracy rate. This low accuracy level indicates that experts may miss a significant number of high-risk and suspicious job advertisements if they rely solely on the model. Consequently, experts cannot depend on the tool until the error rate for these labels is reduced.

### 6.6.4 Dataset Design Implications

The findings of this study provide key insights to develop datasets to detect human trafficking based on job advertisements on job boards.

**Expert Labelling** The labelling process should involve experts with a deep contextual knowledge of the specific country from which the data are sourced. The workshops revealed that experts, particularly the project manager, relied heavily on their contextual knowledge to classify jobs as high-risk or suspicious. Examples of this contextual knowledge included employment patterns in different locations and familiarity with local companies and their practices. The project manager's knowledge of many local companies and their pay structures was particularly helpful in determining whether a high salary was relevant to classifying a job advertisement as suspicious or high risk. Information about these local companies, such as high schools, is not always available through Web search results.

Experts also used indicators discussed in Chapter 5, however, they are less prevalent or obvious in job advertisements on job boards compared to job advertisements found on social media platforms. Therefore, it is recommended that only experts from the countries of focus be involved in the labelling process, rather than crowdsourcing or relying on experts unfamiliar with the specific country, to ensure accuracy and reliability of the data. In addition, experts should have a thorough understanding of the employment industry. This aligns with existing research suggesting that domain expertise significantly improves the quality of labelled data in complex fields (Springbord, 2023; SuperAI, 2023).

**Data Volume** Although the experiments showed good results with limited data, improving the accuracy of high-risk and suspicious labels requires more data, ideally in the tens of thousands. In contrast to social media platforms, job boards such as Website A do not offer a significant number of high-risk and suspicious jobs, making it difficult to collect sufficient examples. Limited data impact the model's ability to accurately identify these labels. The performance of deep learning algorithms

is significantly improved with access to more data, which helps prevent issues such as overfitting (Shen et al., 2023; Malhotra and Godayal, 2018). Therefore, the performance of RoBERTa on these difficult labels will improve significantly with access to more data.

Obtaining larger datasets can be facilitated through partnerships with job boards and online platforms dedicated to machine learning datasets, such as Kaggle. In addition, contacting researchers who have published work involving similar data and collaborating with third-party companies that have access to such datasets can be effective strategies.

**Difficult Labels** The classification of job advertisements under the "suspicious" label presents significant difficulties during the labelling process. Experts found it challenging to definitively categorise certain advertisements as either non-suspicious or high-risk for human trafficking, as reflected in the models' poor performance with this label. To address this issue, future studies may benefit from providing additional examples or further subdividing this category to more accurately capture nuances. For example, the "suspicious" label can be subdivided into "suspicious-human trafficking" and "suspicious-other" to differentiate between job advertisements that are specifically suspicious for human trafficking and those that are suspicious for other reasons. Reducing ambiguity with the label can improve model performance (Malhotra and Godayal, 2018).

**Iterative Labelling Process** Given the complexity of human trafficking cases, which often involves the difficulty of acquiring ground-truth data and dealing with inconsistencies in manual labelling (Toloka, 2023), thorough and repeated labelling processes are crucial to consider, as mentioned in the P3D Framework 3.4. There is a need for a structured approach that involves individual and joint labelling sessions to develop datasets in the field of human trafficking.

Engaging in multiple iterations and workshops can improve the accuracy of the dataset. The experts involved in the labelling process confirmed a substantial improvement in their accuracy levels with each successive iteration and workshop.

The workshops ensured a comprehensive understanding and consistent labelling of job advertisements, as demonstrated by the high levels of agreement with the labels during the last three workshops.

However, iterative labelling can be very time-consuming and resource-intensive as the number of iterations increases (Kotwel, n.d.; Pičuljan, n.d.). This poses a problem when scaling datasets for improved performance. In some cases, as observed in this thesis, more iterations may actually diminish performance, necessitating even more iterations and further resources. The trade-offs of iterative labelling should be carefully considered at the outset to determine whether the potential improvements in accuracy are worth the additional resources.

There have been increasing calls for the use of LLMs for automatic labelling in fields such as human trafficking. In this domain, it is crucial to combine automatic and manual labelling to ensure accuracy and reliability (Toloka, 2023). When resources are limited, it is recommended to prioritise manual labelling for the "high risk of human trafficking" and "suspicious" labels, as these are the most challenging for models to accurately classify. This approach aligns with the literature suggesting that human annotators should selectively review instances after LLMs have performed the initial labelling (Toloka, 2023). Adopting this hybrid approach can enhance both scalability and accuracy, making it a practical solution for real-world applications.

### 6.6.5 Future Technical Implications

Based on the performance discussed above, machine learning has proven to be a viable tool for developing solutions that can automatically detect job advertisements for human trafficking online using AI. The experiments indicate that machine learning can classify job advertisements with a good degree of accuracy and identify suspicious and high-risk jobs that may be missed by experts.

It is recommended that RoBERTa be used as default to classify fraudulent jobs for human trafficking to the great results shown. Its performance is mainly attributed to its advanced training methodology and architecture. RoBERTa has been trained on a vast and diverse dataset, enhancing its capacity to capture rich

language representations (Y. Liu et al., 2019). In contrast to BERT, RoBERTa employs dynamic masking, which enables the model to encounter a wider variety of training instances, and its bidirectional transformer encoder, unlike GPT models, allows it to fully consider context, thus improving its overall performance (Y. Liu et al., 2019). Furthermore, RoBERTa's utilisation of the BPE tokenizer facilitates efficient handling of rare words and subwords, which will be ideal when classifying fraudulent jobs on social media data. As newer and more refined versions of BERT and other PLMs emerge, the academic community and other stakeholders should continue evaluating RoBERTa's performance against these models.

Further research should explore the use of LLMs that allow fine-tuning on specific datasets to detect fraudulent job postings related to human trafficking, given their promising results. In addition, employing these models for additional purposes, such as generating synthetic data to augment datasets, can serve as a valuable tool for combating human trafficking recruitment. LLMs, due to their advanced architecture and training, deliver state-of-the-art performance in natural language generation, which can be instrumental in creating training data for under-represented categories (Minaee et al., 2023; Kalyan, 2023). Synthetic data generation could be particularly beneficial for rare or difficult labels, such as the "suspicious" category, for which LLMs have already exhibited strong performance. Afterward, experts can verify these synthetic data to ensure their accuracy and reliability (Toloka, 2023). However, given the suboptimal performance of LLMs in identifying advertisements related to high risk of human trafficking, it is advisable to rely on other models and expert annotation for this critical category. Future research will focus on identifying methods to improve the performance of this category with LLMs.

Several challenges must be addressed for the effective application of RoBERTa and other deep learning models in real-world scenarios. A significant concern is overfitting, which is a prevalent issue in deep learning models, particularly when trained on limited or non-diverse datasets (Salma and Liu, 2019; Bejani and Ghatee, 2021). A viable approach to mitigating overfitting is to train the model on a more diverse and comprehensive dataset. For example, validating and augmenting a

dataset, such as the 10K dataset or a subset thereof, with input from experts could serve as an effective strategy for enhancing the data available to the models. Another critical factor is ensuring that the necessary computational resources are available. Deep learning models such as RoBERTa are computationally demanding due to their large size and the need for extensive hyperparameter tuning, which is computationally expensive (Roy et al., 2023; Qi and Xu, 2023; N. C. Thompson et al., 2020; Talaei Khoei et al., 2023). As highlighted in Chapter 5, organisations such as Organisation B often face resource limitations, and these constraints must be considered early in the project planning stages to ensure feasibility and success.

### 6.6.6 Recommendations for Policymakers

This papers highlighted potential regulations that could be enacted to help address human trafficking recruitment on online platforms in Country A.

1. There are limited rules and regulations governing online platforms, which poses risks to potential victims (Dorton, 2019; Department of State, 2023). Policymakers should consider the following regulations with regard to online platforms:

    (a) Impose responsibility on online platforms for their role in enabling human trafficking on their platforms while simultaneously safeguarding the interests of legitimate users. This can be achieved by implementing penalties, such as fines or even banning platforms (Department of State, 2023). An example of such a bill is the UK's Online Safety Bill, which mandates that social media platforms actively monitor and remove illicit content, including human trafficking materials. Non-compliance results in substantial financial penalties of up to £18 million, or 10% of their global turnover (GOV.UK, 2023).

    (b) Require technology companies to collaborate with government-approved anti-trafficking entities to devise methods to prevent and eradicate human trafficking on their platforms (Department of State, 2023). This

requirement should specifically target online platforms such as job boards and social media platforms, which have been confirmed to facilitate human trafficking. The partnership must include the exchange of data and technological resources, while simultaneously safeguarding user privacy due to inevitable privacy concerns. Governments should liaise with privacy experts when creating such a bill. Furthermore, previous research suggests that governments should ensure that technology companies only allow verified job listings to be posted on their platforms (Department of State, 2023).

2. Establishing regulations that guarantee adequate financing for anti-trafficking stakeholders is essential. This funding will be instrumental in empowering stakeholders to develop and implement effective strategies to combat human trafficking, as well as to improve their ability to use these strategies efficiently. The Department of State emphasises the need for governments to provide sufficient resources to anti-trafficking stakeholders such as law enforcement agencies to identify fraudulent job offers online (Department of State, 2023).

Policymakers should aim to collaborate with diverse stakeholders, such as anti-trafficking experts and survivors, to help ensure the effectiveness of human trafficking regulations. By adopting this cooperative approach, legislation can better serve the interests of victims and align with the objectives of anti-trafficking initiatives on the ground.

## 6.7 Conclusion

This chapter aimed to evaluate whether machine learning models can be used to identify fraudulent job ads related to human trafficking. To achieve this objective, an expert-labelled dataset was generated and used to train various algorithms, which were classified into three primary approaches: traditional machine learning techniques, pre-trained neural language models, and large language models. The purpose of employing these three approaches was to determine the most effective

method to obtain human trafficking data on fraudulent jobs. RoBERTa was the best-performing model due to its large and diverse training dataset, use of dynamic masking, and bidirectional encoder. Its performance was validated through a workshop with experts, which confirmed satisfactory results. Subsequently, the model labelled 10,000 job advertisements to create a machine-learning labelled dataset.

# 7 Discussion

## Contents


**7.1 Summary of Findings . . . . . 241**
7.1.1 Phase 1: Assessing Anti-trafficking Software . . . . . 243
7.1.2 Phase 2: Analysis of Human Trafficking Recruitment in Africa . . . . . 247
7.1.3 Phase 3: Investigating Fraudulent Jobs for Human Trafficking on Social Media . . . . . 250
7.1.4 Phase 4: A Machine Learning Approach to Detecting Human Trafficking Recruitment . . . . . 254
7.1.5 Reflection on Research Questions . . . . . 256
**7.2 Use of P3D Framework . . . . . 257**
7.2.1 Project Planning . . . . . 257
7.2.2 Data Planning . . . . . 258
7.2.3 Design and Deployment . . . . . 259
7.2.4 Summary . . . . . 259
**7.3 Contributions to Work . . . . . 260**
7.3.1 Empirical Insights . . . . . 260
7.3.2 Design Recommendations and Framework Contributions 261
7.3.3 Technological and Methodological Contributions . . . . 262
**7.4 Limitations and Challenges . . . . . 263**
7.4.1 Phase 1: Assessing Anti-trafficking Software . . . . . 263
7.4.2 Phase 2: Analysis of Human Trafficking Recruitment in Africa . . . . . 264
7.4.3 Phase 3: Investigating Fraudulent Jobs for Human Trafficking on Social Media . . . . . 264
7.4.4 Phase 4: A Machine Learning Approach to Detecting Human Trafficking Recruitment . . . . . 265
**7.5 Challenges of using Computational Methods to Identify and Investigate Recruitment . . . . . 266**

7.5.1 Achieving Accuracy . . . . . 266
7.5.2 Meeting Regulatory Requirements . . . . . 267
7.5.3 Application Integration . . . . . 268
7.5.4 Long Term Performance . . . . . 268
7.5.5 Feasibility of Developing and Using Anti-trafficking Software 270
**7.6 Recommendations for Future Work . . . . . 270**
7.6.1 Broadening the Scope of Research . . . . . 271
7.6.2 Creating Datasets . . . . . 272
7.6.3 Comparing Performance with Social Media Data . . . . 272
7.6.4 Developing and Deploying Real-World Systems . . . . . 272
7.6.5 Developing Optimisation Tools for Anti-trafficking Stakeholders in Africa . . . . . 273
7.6.6 Long-term performance of real-world anti-trafficking tools 274
7.6.7 Studies on Increasing User Trust in Anti-trafficking Software 274
7.6.8 Technical Implications of Anti-trafficking Software . . . 274
7.6.9 Generalising Findings in other African Countries . . . . 275
7.6.10 Refining the P3D Framework . . . . . 275
**7.7 Reflection and Recommendations . . . . . 276**
7.7.1 Academic Community . . . . . 276
7.7.2 Policymakers . . . . . 279
7.7.3 Anti-trafficking Stakeholders . . . . . 280
7.7.4 Final Thoughts . . . . . 281

---

This chapter summarises the main findings and contributions of the thesis, discusses the challenges of using computational methods to detect human trafficking recruitment online, and offers recommendations for future work. In addition, it reflects on the work and provides recommendations for key stakeholders.

## 7.1 Summary of Findings

This thesis aimed to investigate and detect human trafficking on online platforms. A review of the existing literature identified several issues that this research sought to address.

Chapter 2 provided a theoretical background on human trafficking and technology-facilitated trafficking, with a focus on online recruitment. The literature indicates that traffickers are increasingly relying on online platforms to recruit individuals for human trafficking, and these platforms have become the primary method of recruiting victims (OSCE, 2020; Latonero and Movius, 2011; Department of State,

2023; United Nations Office on Drugs and Crime, 2018; Europol, 2014; Ibanez and Suthers, 2014; US Department of State, 2019; Chen et al., 2019). The literature has shown that traffickers prefer online recruitment because it 1) scales easily, allowing them to access victims from diverse geographical areas, 2) reduces the risk of detection by law enforcement, 3) allows the implementation of measures to prevent detection, and 4) facilitates easy access to minors through online grooming, particularly during the Covid-19 pandemic (Progetto Tenda et al., 2021; Anthony, 2018; Sykiotou, 2007; Arronte, 2018).

However, limited research has been conducted to understand the intricacies of online recruitment (Progetto Tenda et al., 2021; Feehs and Wheeler, 2020; United Nations Office on Drugs and Crime, 2020; Thorn, 2018; Gezinski and Gonzalez-Pons, 2022). Most existing research on technology-facilitated trafficking has focused on understanding the exploitation stage, where victims have already been recruited and technology is used to advertise them. There have been calls for more studies on online recruitment, as highlighted in the literature spanning more than a decade up to the present (Department of State, 2023; Thorn, 2018; Greiman and Bain, 2013; Ibanez and Suthers, 2014; Latonero, 2012; Zhou et al., 2024). However, despite calls for further research, there have been limited advancements.

In addition, Chapter 2 also reviewed the existing computational approaches employed to detect human trafficking online, with the majority of efforts focused on identifying trafficking activities during the exploitation stages. Existing datasets for developing technologies have also focused on the exploitation stage, with the majority of the datasets based on escort advertisements (Dubrawski et al., 2015; Alvari et al., 2017; Tong et al., 2017; Alvari et al., 2016; Zhu et al., 2019; Wang et al., 2020; Paula et al., 2023; Esfahani et al., 2020; J. Liu et al., 2023; Summers et al., 2023; Wiriyakun and Kurutach, 2021; Vajiac et al., 2023; Nair et al., 2022; Ardakani, 2020; Freeman et al., 2022). The lack of computational work on online recruitment can be attributed to several factors: 1) Limited datasets on recruitment activities, as most recruitment occurs on social media platforms, making data access more difficult compared to online exploitation data, which is found on public adult

and classified websites; 2) A limited understanding of the methodologies associated with the recruitment phase, with data availability being a contributing factor; and 3) An insufficient understanding of the investigative processes related to recruitment.

Furthermore, the literature review in Chapter 2 highlighted stakeholders' concerns regarding the potential risks and negative consequences associated with the use of anti-trafficking tools (Latonero, 2012; Ediae and Kuteesa, 2024; Deeb-Swihart et al., 2022). A significant aspect of these concerns is the absence of regulations that govern their development and implementation (OSCE, 2020). However, few evaluations of deployed anti-trafficking systems have been conducted to identify these risks and consequences. This lack of evaluation has resulted in a lack of measures to guide the development and research necessary to ensure the responsible outcomes. Failure to identify the potential impact of these tools may pose significant risks to both the victims and society, highlighting the need to address these issues (OSCE, 2020; Latonero and Movius, 2011; Musto et al., 2020).

These identified gaps collectively shaped four research questions. This thesis employed a mixed methods approach to address these questions, structured into four phases. These phases are summarised in Figure 7.1. The following sections outline the research questions, the methods used to address each research question, and discuss the key findings derived from them.

### 7.1.1 Phase 1: Assessing Anti-trafficking Software

The first phase of the thesis aimed to address the first research question with the findings presented in Chapter 3. The research question was: *RQ1: What are the challenges and potential unintended consequences of developing and using anti-trafficking software, and how can these issues be mitigated?* This phase involved conducting interviews with academic, private, and public sector stakeholders in the US and the UK, using the RRI framework as a guide. The initial focus was on the US, as the majority of anti-trafficking instruments have been developed in that country, and it constituted the primary area of investigation. The UK was incorporated for comparative analytical purposes. The findings from this phase

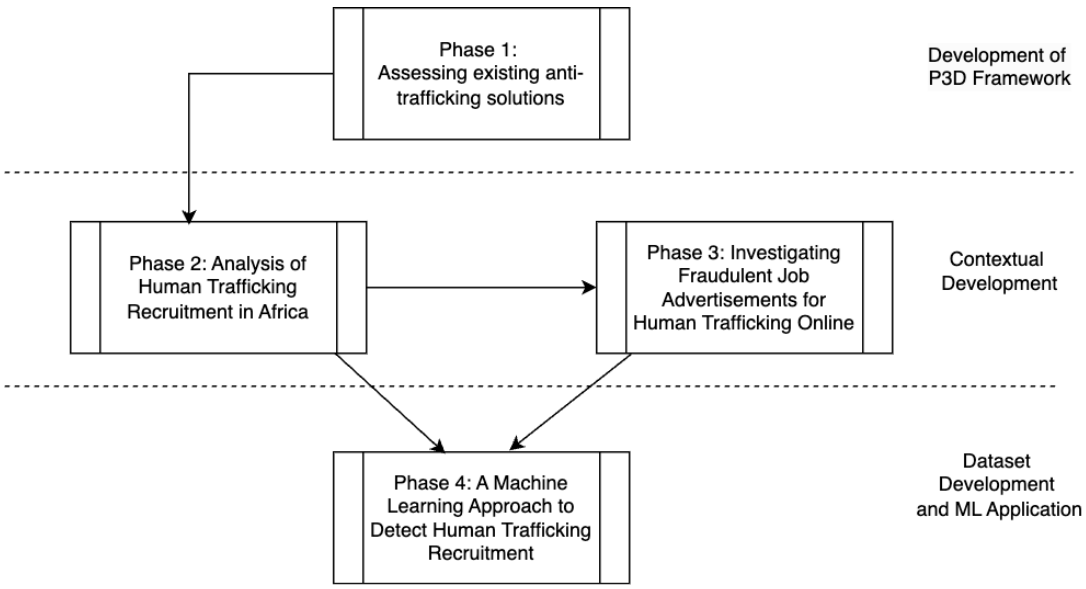


**Figure 7.1:** Summary of Thesis

highlighted the positive and negative outcomes associated with the development of anti-trafficking software and provided recommendations to mitigate negative effects.

The positive consequences confirm that these tools help law enforcement conduct human trafficking investigations, significantly reducing time expenditure and improving efficiency (Marinus Analytics, 2022; Flores, 2022; Pendergrass, 2018). These findings also indicate that the tools facilitate the gathering of intelligence on other criminal activities involving human traffickers, which is advantageous because traffickers can be prosecuted for these crimes, thereby contributing to the prevention of further exploitation. Human traffickers are often convicted of other offences, rather than trafficking itself, because of the complexity of trafficking cases (OSCE, 2020; Ediae and Kuteesa, 2024). Furthermore, it was determined that these tools have facilitated access to critical data, which has contributed to substantial changes such as enhanced decision-making processes and optimised resource allocation in anti-trafficking work.

The identified challenges and negative consequences corroborate the concerns raised in the literature, confirming that risks and negative unintended consequences can arise from the development and use of anti-trafficking software (OSCE, 2020; Latonero and Movius, 2011; Latonero, 2012; Deeb-Swihart et al., 2019; Pendergrass,

2018; Ediae and Kuteesa, 2024; Rivas et al., 2024). These findings extend the privacy and bias concerns discussed in previous studies while also highlighting novel issues that have not been adequately addressed in the existing literature on anti-trafficking tools.

Negative unintended consequences significantly impact three main groups of individuals: victims, vulnerable populations, and misclassified traffickers. Misclassification of victims can lead to the victimisation of sex workers and marginalised communities. Sex workers, in particular, have been substantially affected because much of the data used by these tools originate from their online information, presenting challenges such as potential arrest and invasion of privacy. Furthermore, existing legislation contributes to these adverse effects on sex workers. Since the implementation of the FOSTA-SESTA Act in 2018, numerous sex workers have been banned from online platforms, forcing them to return to street work, leading to increased violence against them, especially workers from marginalised communities (Chamberlain, 2019; Simon, 2018; Marshall, 2016). For example, a sex worker reported that the introduction of FOSTA-SESTA led to 13 missing individuals being reported and two confirmed dead within the first month of the law's enactment (Simon, 2018). As discussed previously, FOSTA-SESTA requires platform providers to be responsible for content related to prostitution (whether consensual or not) posted on their platforms (Romano, 2018), creating incentives for providers to remove suspected sex workers. In addition, misclassification of traffickers can have long-term impacts on individuals, even if they are cleared of charges due to stigmatisation. The potential dangers of misclassification highlight the need to minimise false positives while considering privacy risks.

Furthermore, as more cases are investigated, law enforcement agencies are shutting down websites that facilitate human trafficking. However, the closure of these platforms also negatively affects sex workers who rely heavily on them for their work; the closure of popular platforms has resulted in more sex workers operating on the streets or being controlled by third parties (Erin Taylor, 2019; Tichenor, 2020). For example, a sex worker discusses how the shutdown of Backpage, which

was responsible for 80% of the online sex market, affected their ability to work safely and earn a livelihood, forcing them into more dangerous environments and financial difficulties (Simon, 2018). In addition, traffickers are identifying methods to evade detection by law enforcement, making detection more challenging. In addition, the closure of platforms causes a displacement effect on crime, causing traffickers to spread out to multiple platforms, further complicating detection efforts (OSCE, 2020). Victims may also face an increased risk of violence when using anti-trafficking tools to seek support if traffickers become aware of their efforts to obtain help (Simonson, 2021; Deeb-Swihart, 2022). This highlights the need for a more careful design of technologies targeting victims.

The challenges associated with anti-trafficking tools have revealed issues such as inadequate funding and lack of data, both of which can impact their performance. Funding is a significant challenge in the development of anti-trafficking software (Mitchell and Boyd, 2014; Deeb-Swihart et al., 2019), and efforts should be made to secure more financial support for anti-trafficking stakeholders while identifying cost-effective ways to develop these tools. Another major concern is the lack of data sharing among stakeholders, which affects the development and quality of anti-trafficking efforts (Campana, 2022; Ediae and Kuteesa, 2024; Rivas et al., 2024). Reluctance to share information can be attributed to several factors, such as legal concerns, the need to obtain consent from victims, the responsibility of NGOs to protect victims from negative unintended consequences, the ethical obligations of anti-trafficking stakeholders and the complexity of data sharing legislation (Campana, 2022; Ediae and Kuteesa, 2024). There are recommendations for developing a tool that would allow front-line organisations, including technology companies that combat human trafficking, to share data that could facilitate the identification of traffickers and ensure successful prosecutions (Mitchell and Boyd, 2014). Insights into the challenges and unintended negative consequences of anti-trafficking tools have highlighted the need for more legislation to help mitigate these effects. There is a need to consider whether existing laws and efforts are sufficient to protect vulnerable individuals from the unintended negative consequences of using

these tools, particularly whether there is adequate justification for the continued victimisation of sex workers. Moreover, these challenges and negative consequences have inspired the creation of the P3D Framework, which guides stakeholders in developing anti-trafficking software and informs the subsequent phases of this research to ensure responsible outcomes. This framework addresses concerns raised in previous research regarding the lack of measures that guide the development of anti-trafficking tools and the need to prioritise issues such as privacy and security (OSCE, 2020; Department of State, 2023; Latonero, 2012; Deeb-Swihart et al., 2019; Ediae and Kuteesa, 2024). Discussions with developers have revealed that they are not always aware of the unintended consequences of their technologies, which is why such a framework will be a useful tool during development.

Furthermore, discussions with participants in Chapter 3 and other experts consulted during the initial phases of the thesis led to the recommendation to shift the focus to Africa due to the significant unintended negative consequences observed in Western countries, particularly the US, such as the victimisation of vulnerable individuals and other adverse effects on victims. In addition, these stakeholders expressed concerns about the limited research on technology-facilitated trafficking in Africa and advised a refocus to achieve greater impact. Previous literature has also recommended extending research on online recruitment to regions outside the US (Anthony, 2018). These recommendations led to the shift to Africa, providing an opportunity to maximise the impact of the research by gaining insight into an underexplored area.

### 7.1.2 Phase 2: Analysis of Human Trafficking Recruitment in Africa

The second research question was examined in the second phase of this thesis, with the results presented in Chapter 4. The research question was: *RQ2: How do human traffickers recruit victims both offline and online, and what key factors and patterns contribute to this process?* This question addressed the issue of limited research in this area.

The objective of this phase was to provide contextual information on human trafficking recruitment by examining patterns in different African countries, both offline and online. These insights respond to calls for more research on online recruitment, as highlighted in previous studies (Department of State, 2023; Anthony, 2018; Thorn, 2018; Greiman and Bain, 2013; Ibanez and Suthers, 2014; Latonero, 2012; Zhou et al., 2024). Furthermore, this objective addresses the recommendations from the participants in Chapter 3 in the importance of gaining contextual knowledge of the problem area. To achieve this, data on human trafficking recruitment in 18 African countries were provided by a global organisation, Organisation A. Using data visualisation and statistical analysis, the data were analysed to identify recruitment methods, victim profiles, recruitment processes and recurring trends in human trafficking within the region. Furthermore, the research discusses how the organisation addresses human trafficking and the potential role of technology in these efforts.

The majority of victims in the dataset were youth and children, raising significant concerns about the high rate at which they are recruited in Sub-Saharan Africa. This finding aligns with reports from the United Nations that child trafficking is the most prevalent form of recruitment in Sub-Saharan Africa, with the majority of children recruited for forced labour (United Nations: Office on Drugs and Crime, 2023). A substantial proportion of these children and youth were students, highlighting the need for awareness campaigns in educational institutions. Using both technological and offline resources should be employed to create targeted campaigns for children (Mitchell and Boyd, 2014). Research has shown that involving children in the design of campaigns aimed at them is effective (Horowitz Research, 2023), and this approach should be applied to anti-trafficking efforts.

The findings also revealed that victims are often recruited under the guise of promising opportunities, mainly employment and scholarships. This is consistent with the patterns observed in Western countries, particularly with respect to employment offers, as discussed in Chapter 2 (Anthony, 2018; United Nations Office on Drugs and Crime, 2020). Online recruitment through social media platforms and

job boards plays an important role in Africa, with high salaries and positions in the hospitality and domestic service sectors being common indicators of fraudulent job offers. There are opportunities to develop technologies to verify these scholarships and employment opportunities, as well as empower the public to do the same (Department of State, 2023). Equipping the public to recognise early signs of human trafficking will help prevent potential victims from falling into debt, which often follows recruitment through these methods (Department of State, 2023). Indicators identified in the next phase will also help achieve this goal.

The findings highlighted the work of Organisation A. Their operations allow them to intercept thousands of individuals who would otherwise be recruited for various forms of exploitation, including prostitution and forced labour. This stresses the importance of intercepting victims in transportation hubs, such as airports, where stakeholders can identify them before reaching their final destinations. Expanding their impact and improving operational efficiency are critical. Multidisciplinary solutions are required at these transportation hubs, integrating technology and empowering anti-trafficking stakeholders to detect human trafficking.

This phase highlighted several design implications for Organisation A and other anti-trafficking stakeholders to develop technological solutions. For example, border control systems that leverage advanced image and text analysis capabilities are necessary to increase the number of intercepted cases at transportation hubs. However, it is essential to develop data protection measures to ensure that technologies using image recognition techniques to detect vulnerabilities comply with privacy standards and prevent data misuse (Anderson et al., 2024). These systems should also account for post-deployment modifications through machine learning methodologies, as the environments in which they operate evolve over time (Anthonysamy et al., 2017).

Training front-line personnel at these transportation hubs should also be prioritised. Previous research has indicated that front line personnel, including law enforcement officers, are often not adequately equipped to identify human trafficking cases (Ediae and Kuteesa, 2024). Comprehensive training is imperative for ticket inspectors, airline personnel, immigration officials, security personnel,

and custodial workers to improve their ability to detect trafficking at these hubs. The staff and technological systems must be trained to identify the vulnerabilities indicative of trafficking. As identified in Chapter 4, these vulnerabilities fall into five primary categories: insufficient knowledge of travel arrangements, language and communication barriers, suspicious travel arrangements, vulnerable relationships and dependency, and online or unfamiliar connections.

The findings also show that governments can play a pivotal role in addressing human trafficking recruitment in Africa. The problem is exacerbated by systemic issues such as unemployment, poverty, conflict, and political instability (UNICEF, 2003; Ediae and Kuteesa, 2024; Children International, 2023; Onuoha, 2011; The World Bank, 2022). Governments must address these root causes to mitigate the likelihood of exploitation. Furthermore, protective measures for children in educational institutions are crucial, as the findings show that traffickers frequently target students. Previous literature supports this, noting that traffickers tend to frequent school areas and use relationship-based approaches to recruit victims (Mitchell and Boyd, 2014). Furthermore, the findings and previous research (Department of State, 2023) emphasise the need for government institutions to establish verified organisations that help citizens identify legitimate employment and scholarship opportunities. The following provides details on how fraudulent job opportunities can be identified.

### 7.1.3 Phase 3: Investigating Fraudulent Jobs for Human Trafficking on Social Media

The third phase of this thesis is dedicated to addressing the third research question, the findings of which are presented in Chapter 5. The research question was *RQ3: What are the indicators of fraudulent job advertisements related to online human trafficking recruitment?*

An ethnographic study was conducted that encompassed a combination of in-person and digital observations, interviews, and an analysis of the existing case studies to address the research question. This work was achieved through our

collaboration with Organisation A; however, this phase was limited to a single country, Country A. The team operating in Country A, known as Organisation B, was the only team among those working across 18 African countries that investigated fraudulent job posts on social media. In-person observations involved observing Organisation B's work in investigating fraudulent jobs, while digital observations focused on job advertisement groups on a single social media platform. Interviews were conducted with staff from Organisation B, and the case analysis examined existing cases confirmed by the organisation to be related to human trafficking. This phase facilitated the identification of indicators of fraudulent job advertisements related to online human trafficking recruitment and highlighted the steps involved in investigating such cases. Furthermore, it clarified the information required for the successful execution of these investigations.

Vulnerable populations and job seekers often turn to social media platforms for employment opportunities, many of which operate in unregulated environments (Department of State, 2023). It is difficult to address fraudulent jobs for human trafficking on online platforms because of several factors, including 1) traffickers using multiple distinct social media and job recruitment platforms in different regions, making monitoring and detection difficult; 2) adversarial measures implemented by traffickers to evade detection; 3) the substantial volume of fraudulent job advertisements online; 4) victims' lack of knowledge about human trafficking and fraudulent practices on online platforms, which renders them less cautious when using these platforms; 5) insufficient studies addressing the problem; 6) scarcity of data on the issue; 7) lack of adequate technological solutions; and 8) unregulated online environments (Dorton, 2019; Department of State, 2023; Anthony, 2018; Sykiotou, 2007; Thorn, 2018; Zhou et al., 2024; Campana, 2022; Konrad et al., 2016; Gezinski and Gonzalez-Pons, 2022).

The findings of this phase provide a deeper understanding of how traffickers recruit individuals through social media. The research identified various indicators used by Organisation B to identify and shortlist suspicious advertisements. These indicators can be summarised into three main categories that investigators aim

to detect: 1) incongruencies, 2) irregularities, and 3) familiar misdemeanours. Investigators are continuously looking for incongruencies, pieces of information that contradict or are at odds with industry standards, and contextual understanding of the national job market, such as posts that specify candidates from specific high-risk countries. In recognising these incongruencies, investigators have a deep understanding of the contextual elements relevant to their work. Organisation B ensures that investigators receive routine training in these elements and also encourages them to conduct independent research.

Identifying irregularities involves detecting information that deviates from the norms such as unconventional practices or details that stand out as atypical. Examples include informal writing styles such as the use of emojis in advertisements and inappropriate requests for personal information. Human traffickers tend to use personal information to facilitate their activities. Organisation B detects these irregularities by reviewing advertisements line by line to identify finer details that do not align with the standard job industry practices. Investigators also rely on identifying familiar misdemeanours by detecting the patterns or repeated information used by traffickers. As discussed in Chapter 5, traffickers often reuse information that can aid in detection. Organisation B typically identifies suspicious advertisements by collecting data on confirmed details related to human trafficking, maintaining a record of this information, and comparing the details of new advertisements with these records. Over time, investigators have become well versed in the patterns and details associated with human trafficking, allowing them to detect cases without having to check each against previous records.

There is a prevalent tendency for individuals to place blind trust in job advertisements disseminated through social media platforms. Previous literature (Anthony, 2018) highlights a concerning situation in which migrant workers in the Philippines trusted job advertisements posted on social media more than those of official government regulatory agencies. The US Department of State (Department of State, 2023 has urged social media platforms to collaborate with government agencies and civil society to address this issue. This chapter proposes design considerations that

social media platforms can implement to help reduce the recruitment of victims by traffickers through fraudulent job advertisements.

Organised criminal activity was detected across various account profiles that posted suspicious job advertisements. Examples of organised behaviour include multiple suspicious accounts posting similar advertisements with the same contact details or using similar writing styles. Previous research has noted that the majority of traffickers belong to organised criminal groups, which recruit a larger number of individuals over extended periods and exhibit more violent behaviour during both recruitment and exploitation, highlighting the importance of detecting and removing these traffickers (United Nations: Office on Drugs and Crime, 2023). The US Department of State recommends that technology companies ban accounts linked to human traffickers and implement measures to monitor and mitigate accounts that exhibit suspicious criminal behaviour (Department of State, 2023). Mitigating suspicious accounts can involve restricting their access to features such as job advertisement groups.

Although most studies focus on identifying organised activity and traffickers during the exploitation phase (Arenas et al., 2015; Vajiac et al., 2022, 2023; Nair et al., 2022; Ardakani, 2020), there are opportunities to leverage insights from patterns in suspicious profiles and organised behaviour at an earlier stage. Identifying suspected traffickers can also be achieved by detecting suspicious content posted by traffickers.

Beyond identifying traffickers, analysing the content generated by suspected traffickers can facilitate the identification of trafficker profiles by predicting information such as age and gender. This information is valuable for anti-trafficking stakeholders seeking to enhance their understanding of trafficker demographics. Previous studies have demonstrated the feasibility of using NLP techniques to predict age and gender, even in sparse and noisy data and social media data (Peersman et al., 2022; Peersman, 2018).

A significant challenge observed in this phase is the lack of collaboration between Organisation B and law enforcement agencies to identify fraudulent jobs for human

trafficking. This lack of cooperation has the potential to diminish the impact of Organisation B within the country and may pose substantial risks to the staff. This issue reflects a broader problem in this field, where collaboration presents a considerable challenge, as discussed previously. Although the overarching goal of combating human trafficking is shared, different entities often have divergent goals, which can impede effective cooperation. Identifying areas of common interest in which the outcomes are mutually beneficial is crucial (Freeman et al., 2022).

### 7.1.4 Phase 4: A Machine Learning Approach to Detecting Human Trafficking Recruitment

The final phase of the thesis focused on addressing the fourth research question, with the findings presented in Chapter 6. The research question was: *RQ4: How can a set of models be developed to identify fraudulent job advertisements for online human trafficking recruitment?* This phase aimed to address the gaps in existing research, particularly the minimal use of computational methods to combat online human trafficking recruitment and the absence of state-of-the-art AI techniques for detecting recruitment efforts.

The dataset developed in this chapter was used for experiments using three primary approaches: traditional machine learning methods, PLMs, and LLMs. Traditional approaches were selected because of their appropriateness in the anti-trafficking field, as they yield satisfactory results with limited data and perform efficiently with minimal computational resources (V. Khanna, 2024; Explainable AI, 2024). These factors are significant because of the scarcity of data in the field and the fact that most anti-trafficking stakeholders operate with constrained budgets, which may affect their capacity to use high-computational models. PLMs and LLMs demonstrate excellent performance in classification tasks in various domains, making them suitable for experimentation (Islam et al., 2024; Vaswani et al., 2017; Devlin et al., 2018). Moreover, a limited number of studies have utilised PLMs; however, to the best of our knowledge, LLMs have not yet been employed.

This phase compared the three approaches, evaluated their performance to determine the most suitable model for the dataset, and established a baseline for comparative analysis. The models employed for traditional machine learning approaches and PLMs were based on those used in previous research on detecting human trafficking activity during the exploitation stage. For LLMs, a family of GPT models, recognised as among the best performing LLMs, was adapted for this study (Minaee et al., 2023; Kalyan, 2023). RoBERTa, the most effective model based on accuracy, F1, precision and recall metrics, was selected to label 10,000 unlabelled job advertisements. The results are consistent with previous research, demonstrating high performance in detecting human trafficking activity (J. Liu et al., 2023; A. Rodriguez, n.d.) and in general classification tasks in similar domains (Joy et al., 2022). Consequently, it is recommended that RoBERTa be used as the default model for classifying fraudulent job advertisements related to human trafficking. Its performance is largely attributed to its training on a vast and diverse dataset, dynamic masking strategy, bidirectional encoder, BPE tokenizer (Y. Liu et al., 2019; A. Rodriguez, n.d.).

Before labelling the 10,000 job advertisements, experts evaluated RoBERTa's performance on a separate set of job advertisements, confirming a good level of performance. This new 10K dataset contributes to an expanding corpus of datasets in the field (Tong et al., 2017; Esfahani et al., 2020). Previously developed 10K datasets for similar purposes have shown significant value in academic and industrial contexts, helping stakeholders in efforts to detect human trafficking (Zhu et al., 2019; Wang et al., 2020; Summers et al., 2023; Wiriyakun and Kurutach, 2021; Vajiac et al., 2023). However, it is recommended that the 10K machine-labelled dataset generated in this phase undergo expert validation to enhance the accuracy. Training RoBERTa on more diverse data, such as the validated 10K dataset, will help prevent overfitting, a common issue in deep learning models, particularly when using smaller datasets such as the 1K dataset (Salma and Liu, 2019; Bejani and Ghatee, 2021). Furthermore, training on the 10K dataset will enhance RoBERTa's capacity to more effectively classify "suspicious" and "high risk of human trafficking" labels.

Although data collection methods and analyses varied across phases, the findings of each phase contributed to a deeper understanding and influenced the direction of subsequent phases within the thesis. These phases collectively addressed the four research questions and the general objective of investigating and detecting human trafficking recruitment on online platforms.

### 7.1.5 Reflection on Research Questions

The following section summarises the extent to which each research question was answered based on the findings presented above:

- RQ1 (Challenges and consequences of anti-trafficking software) - Phase 1 identified the challenges, as well as the positive and negative consequences associated with existing anti-trafficking tools, which subsequently informed the development of the P3D Framework.

- RQ2 (Human trafficking recruitment patterns in Africa) - Phase 2 conducted an analysis of trafficking recruitment data from 18 African countries, providing insights into recruitment strategies, recruitment indicators, and recurring trends. This analysis therefore improves our understanding of human trafficking recruitment processes.

- RQ3 (Indicators of fraudulent job advertisements for trafficking recruitment)- Phase 3 revealed various strategies employed by traffickers to recruit victims through fraudulent job advertisements on online platforms, thereby highlighting the associated indicators.

- RQ4 (AI models for detecting fraudulent job ads)- Phase 4 successfully evaluated various machine learning models while taking into account the P3D Framework, thereby establishing a robust benchmark for the detection of trafficking-related job advertisements.

## 7.2 Use of P3D Framework

Chapter 3 introduced the P3D Framework 3.4 as a guide for developing and deploying anti-trafficking software. As previously mentioned, the framework was developed based on the empirical findings and the AREA Framework. The core recommendations of the P3D Framework emerged from themes identified in the participant responses and a review of the literature. Additionally, it incorporates key elements of the AREA Framework, such as reflection, to help minimise the potential harm caused by these tools.

This thesis adapts components of the P3D Framework to develop a dataset and models for detecting fraudulent job postings linked to human trafficking. Integrating the recommendations from the framework has contributed to more responsible outcomes. However, not all components were applied, as this thesis did not involve the development and deployment of a software tool.

The following sections discuss how this thesis integrates the four key components of the framework.

### 7.2.1 Project Planning

The first component of the framework focusses on project planning; it highlights the importance of defining and understanding the problem to be addressed. Project planning was initially explored in Chapter 2, where a review of existing research facilitated the scoping and contextualisation of the problem. The definition of the problem was further refined by engaging stakeholders in Chapter 3, which led to a change in the focus region. Furthermore, data visualisation in Chapter 4 provided additional quantitative insights and narrowed the focus to a specific country.

Furthermore, project planning emphasises the importance of engaging stakeholders in the discussion of the proposed solution. This was accomplished in this thesis by consulting with more than 35 anti-trafficking experts from various countries, such as the UK, the US, Nigeria, and South Africa on the research. This was done through informal meetings or during formal meetings when a study was in progress.

Furthermore, an observational study was conducted with experts in Chapter 5 to gain empirical insight and better contextualise the proposed technical solution.

Another key component of project planning is an ethical review of the proposed solution. Several ethical review processes were carried out at different stages of the thesis (Chapters 3-6) to ensure that the project's methods were in line with best practices and to minimise any negative outcomes. The questions raised during the project planning reflection stage were addressed through the work conducted in Chapters 1-5. All of the work during project planning helped define the objectives of the technical work for this chapter.

### 7.2.2 Data Planning

The second phase of the framework involved data planning, starting with the best practices for data collection and annotation. This thesis collected only relevant data from job advertisements in Country A and did not acquire other types of data. Furthermore, in accordance with the recommendations of the P3D Framework, the annotation process was conducted with the assistance of two experts with relevant experience and knowledge of the country of interest and the job sector.

Addressing bias is another key component of this phase and various measures were taken to reduce it. To mitigate annotation bias, IRR was implemented to ensure that experts achieved a consistent level of agreement, thus reducing the risk of individual bias (Bonnet, 2023). Furthermore, the dataset was developed in three iterations, incorporating individual and group labelling processes to further reduce bias. These iterative and collaborative processes introduced variability, reduced errors and inconsistencies, and minimised individual biases through continuous refinement.

This phase also emphasised the implementation of privacy measures. These included anonymising the names of online platforms, organisations, and participants throughout the thesis. During the initial data collection, the crawling process was designed to avoid collecting personal data. However, the third-party dataset contained company emails and phone numbers, which were removed using Python code. Additionally, the local machine used for storing and processing the dataset,

as well as running traditional machine learning models, was password-protected. Access to the AWS server used to run larger models was restricted to me and my supervisors and was also secured by a password.

The final part of this stage involved manual inspection of the 1K dataset. During this inspection, it was observed that some job descriptions contained contact information. A code was created to automatically remove the presence of this information in the job description column.

### 7.2.3 Design and Deployment

The design phase of the framework required consideration of several factors. Although most were not applied, as they are more relevant during the development of a complete anti-trafficking tool, some were addressed in this thesis. The P3D Framework encourages understanding the location where the tool will be deployed, considering how it might be misused, and how it will serve its users. The potential misuse of the tool was examined in Chapter 3, while Chapter 5 provided information on the environment in which the tool might be deployed and discussed its potential features and functionalities. The P3D Framework also emphasises the participation of stakeholders throughout the development process, which was addressed in this chapter by engaging stakeholders in the data annotation and evaluation processes.

The measures outlined in the deployment phase were not implemented, as they were primarily intended as a fully developed and ready-to-deploy tool.

### 7.2.4 Summary

The use of the P3D Framework in this thesis proved highly valuable in considering various elements that might otherwise be overlooked, which is one of its primary goals. It aims to help developers identify and address areas they may not typically consider while also encouraging them to develop strategies to mitigate harms such as breaches of privacy and bias. Stakeholders can further enhance these efforts by engaging experts in areas such as privacy and bias to develop tailored strategies for their projects.

A key lesson learned from incorporating the framework was the need for sufficient resources, such as funding and time, to achieve objectives such as reducing bias during data annotation. Fully integrating the framework may not always be feasible for PhD researchers or anti-trafficking stakeholders with limited resources. Therefore, stakeholders are encouraged to explore more efficient and customised ways to incorporate its principles into their projects. This thesis demonstrates how some elements can be incorporated into the work without adhering to a rigid systematic approach.

Developers and researchers should always customise the framework to address regional differences, varying technological approaches, and evolving challenges in the fight against human trafficking.

## 7.3 Contributions to Work

The four phases described above produced several contributions that can be classified into three main areas: empirical insights, design recommendations and frameworks, and technological and methodological contributions.

### 7.3.1 Empirical Insights

This thesis produced empirical insights in four main areas.

- **C1** - Exploration of the risks, obstacles, and unintended side effects of anti-trafficking strategies.
- **C2** - An examination of human trafficking recruitment methods in Africa, exploring both online and offline channels, as well as the strategies employed by organisations to combat this issue.
- **C3** - Information on the methods human traffickers employ on social media to recruit individuals, highlighting indicators of suspicious job advertisements and account profiles.

- **C4** - Insights into how anti-trafficking organisations currently investigate and identify instances of fraudulent job advertisements related to human trafficking on online platforms.

These areas have been under-researched, prompting calls for further studies to provide a deeper understanding (Progetto Tenda et al., 2021; Feehs and Wheeler, 2020; United Nations Office on Drugs and Crime, 2020; Thorn, 2018; Greiman and Bain, 2013; Ibanez and Suthers, 2014; Latonero, 2012). The empirical insights gained from this research are valuable to several key groups: **1) the technology community**, where the insights guide the design and development of software tools, helping to create solutions that mitigate negative unintended consequences; **2) the academic community**, as the insights contribute additional knowledge on technology-facilitated trafficking and encourage further research and interdisciplinary collaboration; **and 3) policymakers and anti-trafficking organisations**, who can leverage these insights to inform decision making, shape policy and direct future research efforts.

### 7.3.2 Design Recommendations and Framework Contributions

This thesis produced design recommendations and framework contributions in four main areas:

- **C5** - Created the P3D Framework, which is designed to guide the development and deployment of anti-trafficking software while minimising unintended negative consequences.

- **C6** - Design implications for technology solutions aimed at detecting recruitment activities in Africa.

- **C7** - Design implications for social media companies to consider when designing, developing, and iterating their platforms to improve safety against human trafficking recruitment.

- **C8** - Design implications for creating a dataset focused on human trafficking recruitment.

These design recommendations and framework will help stakeholders in **academic and industry settings** to develop software tools that are contextually relevant, ethically sound, and effective in detecting technology-facilitated trafficking.

### 7.3.3 Technological and Methodological Contributions

The thesis made several technological and methodological contributions. These contributions can be seen in five main areas.

- **C9** -Demonstrated how RRI principles can be used to evaluate existing software with the aim of identifying measures to mitigate negative unintended consequences.
- **C10** - Demonstration of how ethnographic studies can influence the design and development of computational approaches aimed at tackling human trafficking.
- **C11** - Demonstration of how the P3D Framework can be used to inform the development of datasets and computational approaches that address human trafficking.
- **C12** - Creation of an expert-labelled dataset on fraudulent job advertisements for human trafficking.
- **C13** - Benchmarking traditional machine learning models, pre-trained neural language models, and large language models against each other to classify fraudulent job postings related to human trafficking. This provided technical implications for future model development, to improve accuracy in all three approaches.
- **C14** - Deployment of a model that can classify online job postings as suspicious, high-risk for human trafficking, or not suspicious.
- **C15** - Creation of a machine-labelled dataset on fraudulent job advertisements.

These technological and methodological contributions offer several benefits to both the academic community and industry by: **1) providing new ways** to investigate and address technology-facilitated trafficking, combining qualitative and quantitative approaches. This can be adapted in various countries in the Global North and South where such work is lacking; **2) providing datasets** that can help develop and test computational approaches to tackle human trafficking recruitment online; and **3) establishing model results** that **provide baselines** for the academic community to benchmark against when developing tools.

## 7.4 Limitations and Challenges

This section discusses the limitations and challenges encountered during the thesis, structured according to its four phases.

### 7.4.1 Phase 1: Assessing Anti-trafficking Software

The findings from this phase reflect the understanding of 13 participants about Victim/Trafficker Identification tools **in the US and UK**, and **do not cover all types of anti-trafficking tools**. Instead, this study focusses on a better understanding of the impact of these tools in the two countries, based on the experiences and views of the participants. The experts discussed various stakeholder practices; however, practices may vary between specific locations (e.g. countries and cities) and organisations (Freeman et al., 2022).

The P3D Framework also has several limitations. Firstly, it is primarily based on empirical findings from interviews with experts from the United States and the United Kingdom, together with a review of the existing literature. As a result, it reflects the views and experiences of the experts interviewed and the sources reviewed. Furthermore, the framework focusses specifically on tools designed to identify trafficking activity rather than the broader spectrum of anti-trafficking technologies, such as those aimed at victim support. In addition, it places a strong emphasis on data-related aspects, which can overlook other critical factors in anti-trafficking tools.

### 7.4.2 Phase 2: Analysis of Human Trafficking Recruitment in Africa

A significant drawback of this study is that it **does not cover the entire African continent**; instead, it focusses on the countries where the organisation operates. Nigeria, Congo, Ethiopia, Eritrea, and Sudan are among the African nations with a high number of trafficking victims (Global Slavery Index, 2018), however, the datasets only included Ethiopia and not the other countries.

A significant challenge encountered during the study was to **establish collaboration with an organisation**. Despite multiple attempts to engage potential partners, only a few organisations were willing to participate. Furthermore, when a willing partner was found, **a non-disclosure agreement (NDA)** was required before granting access to the data. However, this raised concerns for the department, particularly with regard to the publication of research results. Consequently, they suggested signing a university data-sharing agreement instead of the organisation's NDA. This negotiation process lasted for an estimated three-month period before gaining access to the data, leading to delays in the progress of the PhD research. These limitations and challenges highlight the complexities involved in researching human trafficking, including issues related to data access and collaboration.

### 7.4.3 Phase 3: Investigating Fraudulent Jobs for Human Trafficking on Social Media

The scope of this study was limited to **a single platform and one African country**. However, as previously discussed, human traffickers use multiple online platforms for recruitment and employ fraudulent job advertisements on online platforms in several countries. The challenges in establishing partnerships with other organisations influenced my decision to focus exclusively on Platform A and Country A.

A primary challenge encountered during the study was to **maintain the integrity of the work of Organisation B**. There were concerns about publishing the findings at an academic conference. This challenge is reflective of a broader

issue in this domain. Although research is undeniably important, there is general apprehension about openly sharing investigative methods. The concern stems from the fear that traffickers could adapt and modify their strategies to avoid detection if they become aware of the tactics used to identify them.

Organisation B granted permission for me to **share findings while rigorously anonymising crucial details** such as platform names and country of focus. This approach aimed to prevent inadvertent identification and alleviate concerns. There is a challenge in finding a delicate balance between sharing pivotal insights with the academic community and safeguarding the effectiveness of stakeholders' efforts. Previous research in Chapter 3 highlights the reluctance of anti-trafficking organisations to engage with academia due to these concerns.

### 7.4.4 Phase 4: A Machine Learning Approach to Detecting Human Trafficking Recruitment

This phase had several limitations, including **limited access to data and insufficient resources.** The primary goal of this thesis was to create a dataset based on social media data, as human trafficking recruitment is more prevalent on these online platforms. Unfortunately, due to difficulties in obtaining the data, the thesis used **the second most popular online platform used by traffickers**, based on cases handled by the organisation. Despite efforts to diversify our data by requesting information from other popular job boards, these requests were denied by the respective organisations, highlighting the obstacles in acquiring data and limiting the ability to generalise the work.

Furthermore, **labelling data is expensive** because compensating experts for their time and expertise imposes a significant financial burden (Park et al., 2021; Unterkalmsteiner and Yates, 2023). The cost of labelling increases with each additional session, limiting the number of data points that can be labelled. Consequently, securing funds to support the labelling of a dataset with more than 1,000 examples proved to be a challenging task and limited the scope of the work.

## 7.5 Challenges of using Computational Methods to Identify and Investigate Recruitment

This thesis has demonstrated the potential of using technology to identify and investigate fraudulent job postings related to online human trafficking. However, addressing this complex issue presents several challenges. This section explores some of the challenges associated with the development and deployment of a system designed to detect or investigate such fraudulent jobs. Many of these challenges are also encountered by other anti-trafficking software that addresses different aspects of human trafficking.

### 7.5.1 Achieving Accuracy

Achieving and measuring the accuracy with anti-trafficking software is challenging for several reasons. A significant challenge is the **lack of ground-truth data** in the field. Most solutions rely on expert-labelled data (Simonson, 2021; Tong et al., 2017), which can be prone to errors, affecting the accuracy of the models, as discussed in Chapter 3. This raises concerns about the reliability of these tools, as there is no definitive way to verify the results. Such concerns can also lead to low levels of trust in these tools within society. Experts tend to be highly confident in their labels due to their extensive experience, as evidenced by our work with experts in Organisation B and previous literature (Simonson, 2021; Tong et al., 2017). However, expertise can be subjective, and developers may rely on self-declared expertise rather than experts with hands-on experience or formal qualifications, which could affect accuracy.

Furthermore, the increasing tendency to **adopt a casual writing style** on social media platforms has led to more informal language in job advertisements, which poses challenges for tools that use social media data. The informal language, including slang and context-specific vocabulary related to human trafficking, complicates the accurate classification of job posts by these models. Text from social media is known to influence the accuracy of machine learning models (Alsmadi and Gan, 2019;

Roldós, n.d.). In addition, informal language and slang can vary between countries, further complicating the task of using these tools on an international scale.

**Limited data** availability also affects the accuracy of these tools, as discussed in Chapter 3. The development of machine learning models requires substantial amounts of data; however, organisations are often reluctant to share it. Furthermore, there are challenges in obtaining data related to human trafficking recruitment because of privacy and proprietary concerns. In this thesis, permission to use data scraped from websites was denied. Furthermore, despite reaching out to several companies to request access to such data for research purposes, all of these requests were declined.

### 7.5.2 Meeting Regulatory Requirements

Another challenge in developing anti-trafficking software to detect suspicious posts is the need to comply with regulatory requirements in different regions. **Compliance with these regulations** affects the design of the tool, its overall workflow, and the feasibility of its deployment in multiple countries.

As discussed in Chapter 2, several countries, including Country A, have passed **legislation** that governs data collection and processing. Legal restrictions, including privacy laws and regulations on data crawling, create significant challenges in identifying suspicious job advertisements. For example, the General Data Protection Regulation (GDPR) can limit the collection and processing of data obtained through automated methods such as web crawling, thus hindering the analysis of crucial data related to traffickers and victims (Campana, 2022). A substantial portion of the data related to human trafficking that are used for automation is gathered by crawling the web for advertisements created by traffickers, including those that advertise victims on classified sites or recruit victims through job postings. As a result, these laws create obstacles to data access, as highlighted in this thesis.

Meeting these regulatory requirements can affect the performance of a tool (Campana, 2022). However, these laws are crucial in protecting people from

unintended harm. Therefore, it is essential to strike a balance between safeguarding individual privacy and effectively detecting human trafficking.

### 7.5.3 Application Integration

Developing a tool to assist stakeholders, such as Organisation B, in the investigation process requires access to various types of data, necessitating collaboration. In some cases, this may involve **accessing data from other anti-trafficking stakeholders** or working together to ensure the successful development and deployment of the tool. However, as previously discussed in Chapter 3 and in the section above, fostering collaboration among anti-trafficking stakeholders is a significant challenge in the field (Campana, 2022; Onuoha, 2011).

In addition, **data from general companies** may also be required. For example, developers might need to establish partnerships with various third-party applications to access company information and phone data, which are crucial when developing a system to support stakeholders in investigating trafficking cases. Persuading the general public and private organisations to provide data or integrate their modules with the tool can be difficult, even when the information is publicly available. The lack of collaboration with both anti-trafficking stakeholders and the general public presents substantial challenges to the successful deployment of the tool.

An alternative approach would be for stakeholders to create their own databases containing the information. However, phone numbers are considered personal information and may be subject to privacy laws and ethical considerations.

### 7.5.4 Long Term Performance

The use of AI in developing anti-trafficking software can have long-term impacts on the performance of these systems, potentially disadvantaging stakeholders in their efforts to detect and investigate human trafficking. Several factors can influence performance over time. This section discusses two such factors: **feedback loops and concept drift**.

Feedback loops, while beneficial for tool performance, can also introduce significant risks (Soni, 2022). Risky feedback loops can **influence the types of cases handled by anti-trafficking stakeholders**, such as Organisation B, after the tool is deployed. These loops occur when an AI system's predictions affect the data it subsequently receives, potentially reinforcing biases present in the training data (Soni, 2022; Waters, 2023). For example, if an AI system is biased towards detecting certain types of trafficking cases based on its initial training data, it might overlook other types of cases. The cases it flags and processes can then skew its future training data, particularly if the outcomes are used to retrain the model. This can create **a self-reinforcing cycle**, where the AI system is increasingly focused on a narrow subset of cases, neglecting others.

Concept drift occurs when the statistical properties of the target variable **change over time in unexpected ways**, altering the relationship between the input features and the target variable (Gama et al., 2014; Seldon, 2021; Iguazio, n.d.). As a result, the patterns that the model learnt from historical data may **no longer be relevant to new data**. The tendency of traffickers to adapt their methods, such as using emojis or altering their language, increases the likelihood of concept drift. For example, in recent years, traffickers have started using emojis to communicate illicit activities, bypassing keyword detection systems (Whitney, 2017; Rivas et al., 2024). This introduces **new patterns and features** that the AI system may not have been trained to recognise. Consequently, the model's performance may degrade over time as it struggles to adapt to these changes, leading to concept drift.

Although these are not the only factors that can compromise performance, they are significant issues that prompt stakeholders to consider other technical challenges affecting AI systems. The consequences of concept drift and feedback loops can be severe, impacting potential victims and society both directly and indirectly.

### 7.5.5 Feasibility of Developing and Using Anti-trafficking Software

Developing anti-trafficking tools requires significant resources, including skilled personnel, equipment, and software, all of which incur substantial costs (Campana, 2022). Beyond initial development and implementation, ongoing funding is essential to ensure that the system and its data are regularly updated and monitored (Mitchell and Boyd, 2014). This is crucial for maintaining accuracy and addressing potential performance issues such as concept drift and evolving requirements.

Moreover, the exponential rate at which online data is generated presents additional challenges. Tools designed to process this data for detecting human trafficking must have substantial computational resources to manage large volumes of information.

In addition, traffickers often migrate to new platforms to evade detection or expand their criminal networks. As a result, anti-trafficking tools must be capable of handling data across multiple platforms. However, accessing these data and customising features to meet the specific needs of each platform demands both significant computational and non-computational resources. These requirements make the successful implementation and maintenance of such tools resource intensive, and limited resources can hinder the ability to scale operations effectively.

Many anti-trafficking stakeholders, such as Organisation A and law enforcement agencies, often operate under resource constraints, as reported by participants in Chapters 3 and 4, as well as in previous literature (Mitchell and Boyd, 2014; Ediae and Kuteesa, 2024). These limitations can hinder the development and maintenance of tools, potentially compromising their quality and leading to unfavourable consequences.

## 7.6 Recommendations for Future Work

This thesis explored the challenges of using computational approaches in the preceding section and detailed various challenges and limitations in Sections 7.4.1,

7.4.2, 7.4.3, and 7.4.4. These challenges and limitations highlight opportunities for future research, which are discussed in this section.

### 7.6.1 Broadening the Scope of Research

There is a need to broaden the scope of this research in several ways. The first approach is to conduct further studies that evaluate the effectiveness of other types of anti-trafficking tools beyond Victim/Trafficker Identification systems, and to extend the analysis to additional countries to further enhance the P3D Framework.

The second approach is for researchers to expand this work to countries outside of Africa, where online recruitment is prevalent, such as the US and Italy (Progetto Tenda et al., 2021; Feehs and Wheeler, 2020; United Nations Office on Drugs and Crime, 2020). Expanding these studies will help: 1) better understand recruitment processes in these countries and identify indicators that may differ from those in the African context; 2) gain a comprehensive understanding of this phenomenon on a global scale, which can help stakeholders develop effective strategies to combat this crime (OSCE, 2020; Thorn, 2018; Ibanez and Suthers, 2014); 3) facilitate comparisons of how online recruitment varies across different regions, influenced by factors such as culture and legislation, thus helping to identify both similarities and differences between these areas; and 4) replicate the process of creating labelled datasets and experimenting with various AI methods on these datasets in different countries. Expansion to countries outside of Africa may present challenges in accessing data, similar to the issues encountered in this thesis. Establishing collaborations with stakeholders such as Polaris, OSCE, Europol, and Thorn could potentially help facilitate similar work in the United States and Europe.

A third way to extend the scope is to incorporate more online platforms. Although this study concentrated on fraudulent job postings on a single social media site, it is crucial that the academic community extend its investigation to include a broader range of online platforms. A better understanding of online human trafficking activities can be obtained by researching on multiple platforms and determining the relationships to what happens offline.

### 7.6.2 Creating Datasets

This thesis highlights the need for datasets to facilitate additional research and the development of technological interventions. Although the development of technical solutions and conducting research to combat human trafficking are crucial, a greater emphasis should be placed on creating high-quality labelled datasets.

The creation of datasets should include data collected from online sources, including websites and social media platforms, as well as offline data obtained through fieldwork, such as the datasets described in Chapter 4. The datasets should cover various stages and forms of human trafficking, focussing particularly on under-researched areas like recruitment and different types of exploitation, such as forced labour. The key considerations for creating labelled datasets are detailed in 6.6.4.

### 7.6.3 Comparing Performance with Social Media Data

As previously discussed, there are notable differences between job advertisements posted on job boards and those posted on social media platforms. For example, social media data are often characterised by greater noise and may contain irrelevant or misleading information that can adversely affect the accuracy of the results. These distinctions can affect the effectiveness of the various models and approaches. Models such as Random Forest, with their ensemble nature, and neural networks using regularization techniques, can effectively handle noisy data (Breiman, 2001; M, 2020; RITHP, 2023; DeAngelo et al., 2019; Yiu, 2019). Future research should evaluate the performance of these models on social media data and compare the results with those obtained from job boards in this thesis.

### 7.6.4 Developing and Deploying Real-World Systems

This thesis has demonstrated the potential of state-of-the-art AI models to identify fraudulent job advertisements linked to human trafficking. To maximise their impact in the real world, functional systems should be developed and deployed, such as law enforcement tools and NGO monitoring systems.

For real-world applications, these systems should be capable of processing both text and images, as traffickers often use a combination of these data types to manipulate victims and evade detection, as previously discussed. Previous research (Tong et al., 2017) found that models that incorporate text and images to classify sex trafficking advertisements during the exploitation stage outperformed methods relying solely on one data type. Therefore, this multimodal approach should be investigated for classifying fraudulent job advertisements related to human trafficking. The effectiveness of these adaptations can then be evaluated against the accuracy scores presented in this thesis.

### 7.6.5 Developing Optimisation Tools for Anti-trafficking Stakeholders in Africa

The participants from Chapter 3, the analysis of African countries in Chapter 4, and the ethnographic work in Country A in Chapter 5 emphasise the need for more automated solutions in Africa. It is crucial to have tools that can help anti-trafficking stakeholders in Africa combat human trafficking not only during the recruitment stage but also during the exploitation stage. More work is needed to adapt existing impactful tools, such as those summarised in Table **??**, to African contexts, as well as to develop new tools specifically designed to address the unique challenges of human trafficking on the continent.

For organisations A and B, various tools can be developed to automate processes such as identifying vulnerabilities, managing cases during investigations, and promoting awareness. The suggested design considerations for tools aimed at optimising their work are provided in Sections 4.7.3 and 5.4.3.

Moreover, the findings presented in Chapter 5 have the potential to inform the creation of a rule-based system designed to identify fraudulent job offers associated with human trafficking in Country A. However, it is crucial to ensure that the indicators used are quantifiable and that this system functions solely as an initial support tool for further investigation rather than as a replacement for human expertise.

### 7.6.6 Long-term performance of real-world anti-trafficking tools

Researchers are recommended to identify strategies that should be implemented to address long-term performance issues of anti-trafficking tools such as concept drift and feedback loops. An example approach to mitigate these challenges is through human-in-the-loop methodologies, in which experts are integrated into the decision-making process and models can be adjusted in real time. This methodology enables experts to rectify erroneous predictions or to identify new patterns and potential biases. Furthermore, frequent retraining of the model will ensure that it assimilates and adapts to evolving patterns, thus mitigating the reinforcement of biases.

Measures addressing long-term performance will be important to ensure the relevance of the tool and to avoid various negative consequences that may arise with the use of the tool.

### 7.6.7 Studies on Increasing User Trust in Anti-trafficking Software

Several anti-trafficking stakeholders maintain a significant degree of scepticism about the use of technology to combat human trafficking. As discussed in Chapter 3, some participants expressed concerns about the accuracy of the tools and the potential negative consequences associated with their implementation. This scepticism may impede potential collaboration with other stakeholders, including members of the academic community, who seek to address the issue through technological means. It is imperative to conduct research on methods to improve users' trust in these tools. A key question that warrants consideration is: "What design features can be incorporated into anti-trafficking software to foster greater trust among anti-trafficking stakeholders to increase the likelihood of collaboration?"

### 7.6.8 Technical Implications of Anti-trafficking Software

Further investigation is recommended to evaluate the long-term technical consequences of anti-trafficking software. Although some of the technical consequences

of anti-trafficking software have been identified through this research, there is a need for a more comprehensive study, similar to that presented in Chapter 3, to fully assess the implications of these tools. This study should focus on the technical consequences that impact long-term performance and provide guidance on how to address these challenges in the context of human trafficking.

### 7.6.9 Generalising Findings in other African Countries

The results of Chapter 6 focus on a single country; however, as highlighted in Chapter 4, traffickers also exploit false job offers to recruit victims in other African nations. It is crucial to investigate whether the conclusions drawn in Chapters 5 and 6 can be extended to other Sub-Saharan African countries, where fraudulent job offers for human trafficking also pose significant challenges. The diverse cultures and varying socioeconomic conditions in African countries may have different impacts on trafficking patterns. Furthermore, legislative frameworks can also influence outcomes, as some countries like Kenya and South Africa have enacted laws that address online harms such as human trafficking, while most others have not (National Computer and Cybercrime Coordination Committee, n.d.; Gazette and Notice, 2022).

In addition, it is essential to evaluate the relevance of the insights obtained on human trafficking recruitment in Chapter 4 to other countries that are not included in the datasets but have high levels of human trafficking. Countries such as Nigeria, Congo, Ethiopia, Eritrea and Sudan are African nations with the highest number of trafficking victims (Global Slavery Index, 2018). However, with the exception of Ethiopia, they were not represented in the datasets analysed in Chapter 4. More research should be conducted to confirm that these findings are applicable to these countries, as this could improve overall prevention strategies in Africa.

### 7.6.10 Refining the P3D Framework

Further research is needed to refine the framework into a more generalised version that can be applied across a wider range of tools and national contexts. This should include integrating information from a variety of sources beyond interviews

and broadening the geographical scope of the study. In addition, the framework should be expanded to address currently underexplored aspects, such as algorithm performance. Some recommendations discussed, including the investigation of long-term performance, can also contribute to its refinement.

## 7.7 Reflection and Recommendations

The exploration of the complex nature of human trafficking in this thesis has provided insights that may be valuable to various stakeholders in their efforts to address this issue. This section offers recommendations to the academic community, policymakers, and various anti-trafficking stakeholders based on reflections on the work conducted in this thesis.

### 7.7.1 Academic Community

These considerations aim to assist the academic community in conducting research in this field or in developing technological interventions to address human trafficking.

1. Identifying key anti-trafficking stakeholders at the outset of the project is crucial to ensure successful implementation and avoid delays. However, as previously noted, establishing effective collaboration with anti-trafficking experts remains a challenge. To overcome difficulties in identifying stakeholders, universities can play a pivotal role by forging partnerships with anti-trafficking organisations to foster collaboration, such as offering advanced technical interventions or insights to maximise their efficiency. For example, Polaris collaborated with Harvard University to obtain insights regarding sex trafficking patterns in the United States (Ryan et al., 2016), while the Traffic Analysis Hub engages in partnerships with institutions such as Ohio University and the University of Alabama to exchange insights and data to combat human trafficking (Trafficking Analysis Hub, n.d.). Engaging with key stakeholders was crucial to gaining a deeper understanding of the problem area, refining the objectives of the thesis and conducting research that yielded

meaningful results for their efforts. Researchers should actively identify ways to involve survivors of human trafficking. For example, 1) resources should be allocated and designated to support victims in the event of any complications arising during the engagement process, and 2) liaise and obtain approval from anti-trafficking stakeholders when determining methods of engagement and formulating research questions.

2. Researchers are strongly encouraged to obtain permission from companies like social media platforms and job boards to use their data at the beginning to avoid delays. These entities often take a long time to respond to meeting requests or provide feedback on the use of their data. Researchers can identify incentives such as supporting their research to increase chances of collaboration.

3. When considering the development of datasets and models to identify fraudulent jobs related to human trafficking, it is advisable to prioritise social media data over job boards to avoid delays in data labelling. Social media data are typically easier to label and can provide adequate data for all categories, including "suspicious", "not suspicious", and "high risk of human trafficking". In contrast, labelling advertisements on job boards is more difficult because of the tendency to conform to the standards of these websites. The differences were observed through the involvement of Organization B while investigating social media posts in Chapter 5, and during the data labelling process on job advertisements in Chapter 6." If time or resource constraints are an issue, then focussing on social media data may be the most effective approach.

4. When developing technical solutions to combat human trafficking, it is advisable to conduct preliminary studies to acquire contextual information on the subject matter. By employing a mixed-methods approach, one can gain insights into general trends through quantitative data, while also capturing personal experiences and nuances from anti-trafficking stakeholders using qualitative research methods. This contextual knowledge is crucial for

identifying insights that inform the design of the tool and the methods used in its development. In this thesis, this approach significantly enhances the design and direction of the project. Furthermore, the participants in Chapter 3 emphasised the importance of this approach, as human trafficking is a complex field that requires a thorough understanding for technical solutions to be effective and produce responsible outcomes for its intended users.

5. Researchers are advised to incorporate the P3D Framework into their research studies. The framework should be adapted from the start to help prepare for the research. This framework enabled me to evaluate a range of factors that I might not have otherwise considered. In the technology sector, it is common for professionals to overlook the unintended consequences of their work, as discussed in Chapter 3, which can result in numerous issues. Hence, it is crucial to thoroughly assess the potential consequences and account for both the short- and long-term impacts of work.

The academic community is encouraged to actively engage in technology-facilitated trafficking and contribute toward developing solutions that address this problem, which has a profound impact on millions of individuals. Despite its importance, this area is still in its early stages, and many universities have not yet established dedicated teams to address this problem. It is recommended that academic institutions strengthen efforts to secure funding for research and tool development to combat trafficking, in order to encourage the participation of both senior and junior researchers in these initiatives.

These teams could be specifically dedicated to addressing human trafficking or be integrated into broader groups focused on tackling online harm. This is an opportune time to form such teams, as more countries are introducing regulations to combat online harm, including technology-facilitated trafficking, as discussed in Chapter 2. However, as highlighted in Chapter 3, working on human trafficking can take a toll on mental health because of exposure to explicit and potentially distressing content, especially when dealing with sex trafficking advertisements.

Therefore, these teams could also play a crucial role in providing support systems for researchers, such as access to university counselling services. In addition, support groups for people working in sensitive areas should be established to offer further emotional and psychological support.

Furthermore, universities are encouraged to implement appropriate measures to support fieldwork in sensitive and high-risk areas such as human trafficking research. Ethics departments should collaborate closely with researchers to establish check-in and escalation protocols, ensuring their safety in the field, as unanticipated risks may arise. For example, implementing multiple check-in methods can enhance researcher safety.

### 7.7.2 Policymakers

This thesis highlighted potential regulations that could be enacted to help address human trafficking.

1. There are limited rules and regulations governing social media platforms, which poses risks to potential victims (Dorton, 2019; Department of State, 2023). Policymakers should consider the following regulations with regard to social media companies:

    (a) Impose responsibility on social media companies for their role in enabling human trafficking on their platforms while simultaneously safeguarding the interests of legitimate users. This can be achieved by implementing penalties, such as fines or even banning platforms (Department of State, 2023). An example of such a bill is the UK's Online Safety Bill, which mandates that social media platforms actively monitor and remove illicit content, including human trafficking materials. Non-compliance results in substantial financial penalties of up to £18 million, or 10% of their global turnover (GOV.UK, 2023).

    (b) Permit law enforcement agencies to conduct investigations on social media sites while safeguarding privacy. The law enforcement officials in

Chapter 3 indicated that they face challenges when investigating human cases on these platforms.

(c) Require technology companies to collaborate with government-approved anti-trafficking entities to devise methods to prevent and eradicate human trafficking on their platforms (Department of State, 2023). This requirement should specifically target social media platforms, which have been confirmed to facilitate human trafficking. The partnership must include the exchange of data and technological resources, while simultaneously safeguarding user privacy due to inevitable privacy concerns. Governments should liaise with privacy experts when creating such a bill. Furthermore, previous research suggests that governments should ensure that technology companies only allow verified job listings to be posted on their platforms (Department of State, 2023).

2. Establishing regulations that guarantee adequate financing for anti-trafficking stakeholders is essential. This funding will be instrumental in empowering stakeholders to develop and implement effective strategies to combat human trafficking, as well as to improve their ability to use these strategies efficiently. The Department of State emphasises the need for governments to provide sufficient resources to law enforcement agencies to identify fraudulent job offers online (Department of State, 2023).

Policymakers should aim to collaborate with diverse stakeholders, such as anti-trafficking experts and survivors, to help ensure the effectiveness of human trafficking regulations. By adopting this cooperative approach, legislation can better serve the interests of victims and align with the objectives of anti-trafficking initiatives on the ground.

### 7.7.3 Anti-trafficking Stakeholders

The following are recommendations for anti-trafficking stakeholders.

1. This thesis highlights the importance of collaboration between different stakeholders. It is essential to establish more working groups that encourage collaboration among anti-trafficking stakeholders, as these groups are vital in tackling trafficking problems that require access to various types of data. These working groups should have partnership agreements with formal guidelines for all parties to address privacy and security concerns. In addition, they should devise guidelines on how they can contribute to the research and interventions conducted by external parties.

2. Anti-trafficking stakeholders are encouraged to support researchers and software developers working on anti-trafficking initiatives, whether they engage independently or as part of an organisation. Real-world experience is invaluable in this field and provides unique insights that may not be readily available in the literature.

3. It is crucial that stakeholders remain informed about the advancements in technology employed by human traffickers to expand their operations and increase the number of victims. This knowledge enables stakeholders to ethically employ ways to counter these efforts. Although some stakeholders may not currently use technology in their work, understanding these areas can help identify new methods to optimise their efforts.

4. Survivors should be assigned more specific roles to ensure that their perspectives are adequately represented when participating in projects (Deeb-Swihart, 2022). The input from survivors provides first-hand experiences on issues such as potential harm to victims. They should also implement measures that protect themselves and allow them the freedom to withdraw from the project if complications arise.

### 7.7.4 Final Thoughts

The purpose of this thesis was to examine and identify human trafficking recruitment on online platforms using a mixed-method approach, while adhering to the principles

of responsible research and innovation. Human traffickers exploit a variety of platforms to access and recruit thousands of individuals. Although much work remains to be done to understand and detect human trafficking recruitment on online platforms, this study represents a crucial initial step in that direction. The findings have improved our understanding of recruitment practices and revealed opportunities to leverage technological advancements to identify such activities. Although technology plays a crucial role in addressing human trafficking, stakeholders, including those in academia, must carefully consider its potential negative impacts on victims and society. Frameworks like RRI can help minimise harm, but as demonstrated in this thesis, RRI can also serve as a tool for developing custom frameworks such as the P3D Framework. Stakeholders that leverage technology in this field are encouraged to integrate existing frameworks or develop custom ethical frameworks to maximise the positive impact of their tools while mitigating potential risks.

The issue of human trafficking recruitment on online platforms cannot be resolved by any single stakeholder. It requires the combined efforts of the public sector, private sector, anti-trafficking organisations, and academia. Multidisciplinary solutions are necessary, incorporating both quantitative and qualitative methods, to address this problem. However, stakeholders must address the systemic issues that contribute to the high prevalence of trafficking victims.

# A

# Chapter 3 Supplementary Materials

## A.1 Interview Questions

### A.1.1 Developers

**Role**

- What is the nature of your role and your day-to-day activities?

**Consequences of Developing Anti-trafficking Technologies**

- Describe your experiences in developing anti-trafficking tools that identify sex trafficking activity?
  - What are some of the challenges you experienced while developing the tool?
- What were the effects of the tools on society and law enforcement?
  - What are some of the issues that resulted from the use of the tool that were not intended during the development stage?
- What are the principles and guidelines you employed whilst developing the tools above?

**Recommendation**

- What strategies can be employed to mitigate the challenges and unintended consequences identified in the above section?

**Future Work**

- What are your thoughts on the efficiency of existing tools in curbing online sex trafficking?
- Based on your experience in developing anti-trafficking tools, do you have any recommendations on how the existing technologies can be improved to ensure they are efficient in identifying sex trafficking activity online?

### A.1.2 Law Enforcement

**Role**

- What is the nature of your role and your day-to-day activities?

**Consequences of using Anti-trafficking Technologies**

- What are your experiences of using technologies that identify sex trafficking activity?
  - What are some of the challenges you experienced while using these technologies?
- What are the effects of such technologies on society?
  - What are some of the unintended consequences that resulted from the use of these technologies?

**Recommendations**

- What strategies can be employed to mitigate the challenges and unintended consequences identified in the above section?

**Future Work**

- What are your thoughts on the efficiency of existing technologies in assisting law enforcement with identifying sex trafficking activity?
  - Are there any existing limitations within the existing technologies?
- How can existing anti-trafficking technologies be improved to ensure that they are efficient at identifying sex trafficking activity online?
- What features and technologies would you recommend researchers develop to help identify sex trafficking at its various stages?

### A.1.3 Anti-trafficking Experts

**Role**

- What is the nature of your role and your day-to-day activities?

**Consequences of developing technologies that identify sex trafficking**

- What are your experiences with technologies that identify sex trafficking activity online?
  - What are some of the challenges you are aware of that result from the use of such technologies?
  - What role do these technologies play in ending sex trafficking?
- What are the effects of these technologies that identify sex trafficking on society?
  - What are some of the unintended consequences that arise from the use of these technologies?
- What are some of the guidelines and measures that you would suggest to individuals to take into account before developing technologies that identify sex trafficking?

**Recommendations**

- What are the strategies that can be employed to mitigate the challenges and unintended consequences identified in the section above?

**Future work**

- What are your thoughts on the efficiency of existing technologies in curbing sex trafficking
- What are your recommendations for improving existing technologies to ensure they efficiently identify sex trafficking activity?
- What technologies can be developed to assist your efforts to combat sex trafficking?

## A.2 Consent Form

DPhil project: Online interview study (Human Trafficking professionals) information sheet, February 2022`
Department of Computer Science. University of Oxford
Researcher: Towera Moyo. Towera.moyo@new.ox.ac.uk
PI: Prof. Marina Jirotka. marina.jirotka@cs.ox.ac.uk `

# PARTICIPANT CONSENT FORM

**Computer Science Departmental Research Ethics Committee reference CS_C1A_20_006_001**

**I am conducting online interviews as part of a PhD study titled 'Investigating sex trafficking recruitment on online platforms'. I will hold interviews with a range of experts to identify the challenges and unintended consequences of using anti-trafficking technologies to identify sex trafficking activity, and the strategies that can be used to mitigate them. The interviews will be recorded and transcribed.**

**Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research. You can contact the researcher Towera Moyo, towera.moyo@cs.ox.ac.uk.**

| | | Please tick |
|---|---|---|
| 1. | I confirm that I have read and understood the information sheet dated February 2022 for the above study. I have had the opportunity to consider the information and to ask the researcher questions about my participation. | ☐ |
| 2 | I understand that this study has been reviewed by, and received ethics clearance through, the University of Oxford Computer Science Departmental Research Ethics Committee. | ☐ |
| 3 | I understand that by participating in the study I will be invited to take part in an online interview and that the interview will be recorded. | ☐ |
| 4 | I understand that I can choose whether to join the interview with audio only or with video, and that if I use video I can blur or change my background if I want to. | ☐ |
| 5 | I understand that my participation is voluntary and that I am free to withdraw from taking part in any or all of the data collection activities without giving any reason. Furthermore, I understand that I will be able to request the withdrawal of my data at any time up to two weeks after data collection. | ☐ |
| 6 | I understand that members of project (researcher and supervisors) will have access to the data I provide and that all the data will be stored securely in keeping with the regulations of the University of Oxford. | ☐ |
| 7 | I understand that the findings of the study will contribute to the researcher's DPhil thesis and that portions of my interview might be used in this work. | ☐ |

## A.3 Images - Codes and Themes

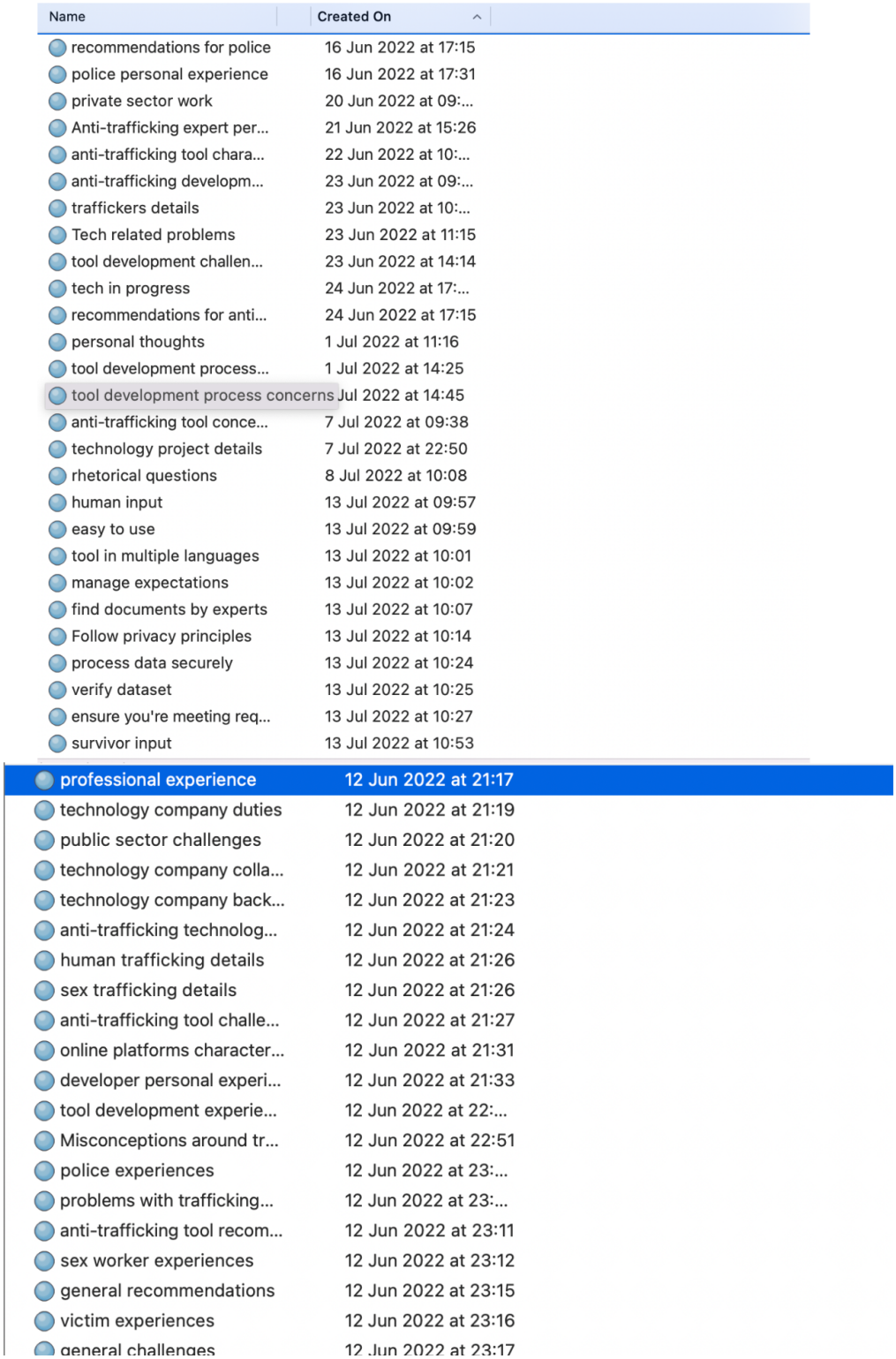


**Figure A.1:** First Level Coding

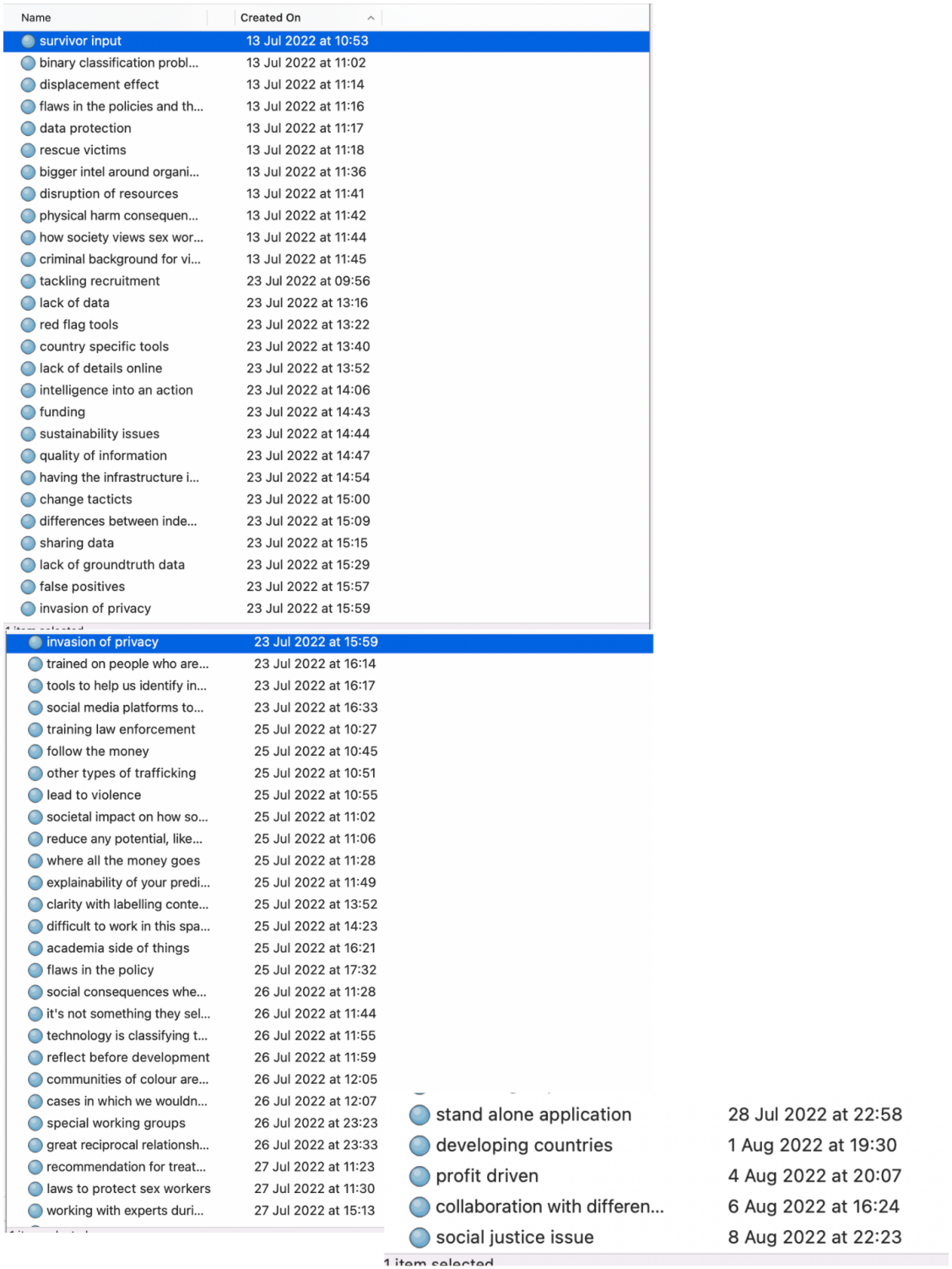


**Figure A.2:** First Level Coding

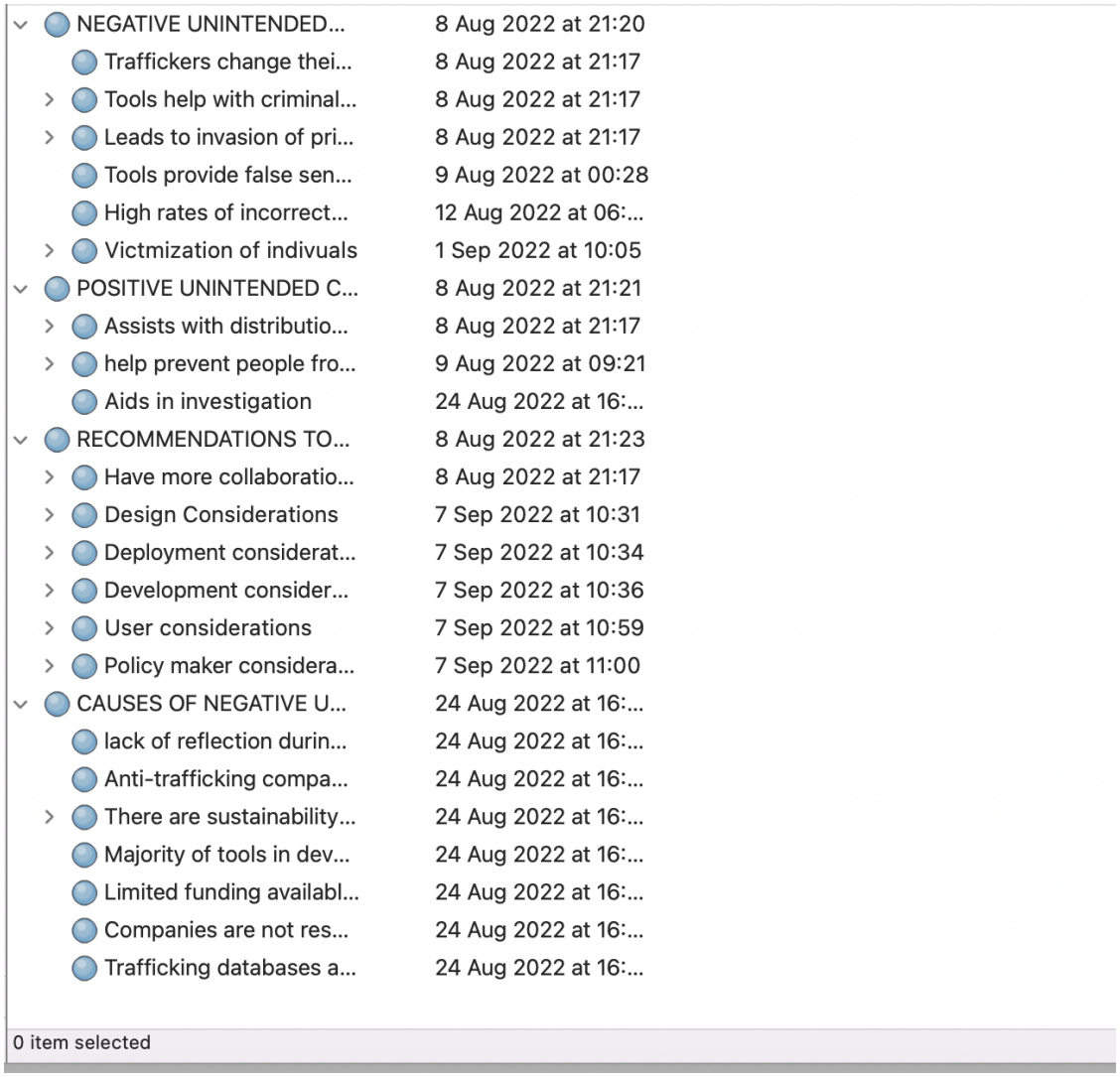


**Figure A.3:** Chapter 4 - Final Level Coding

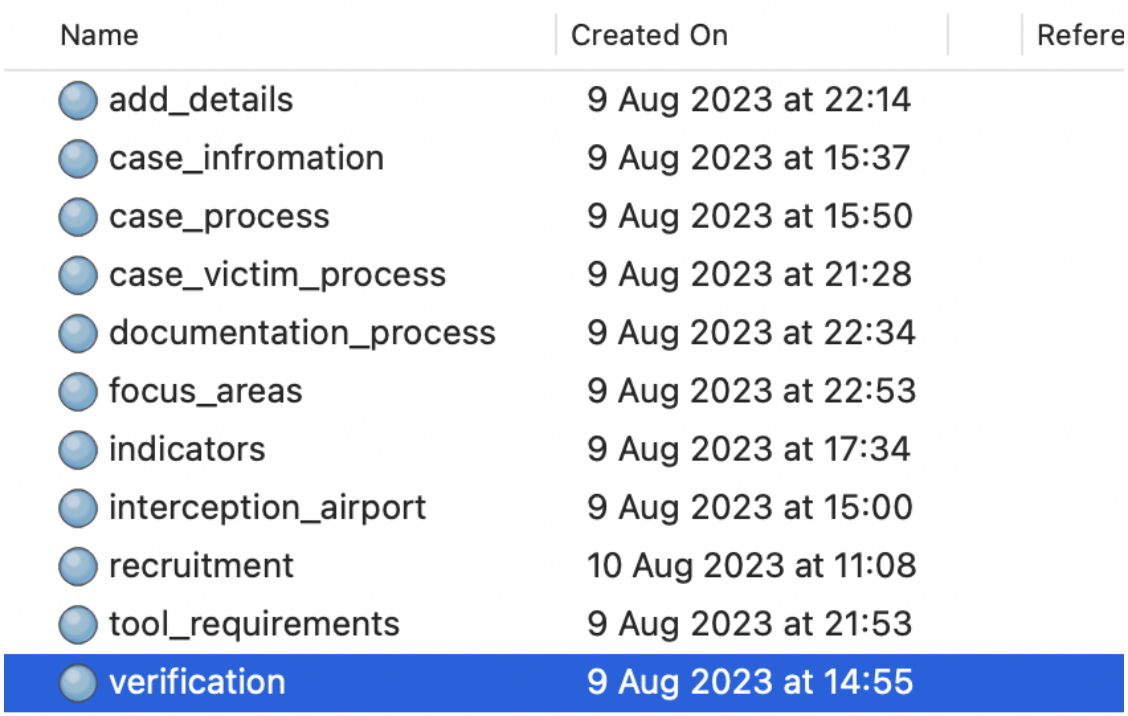


**Figure A.4:** Study 6 - Coding for Observations and Interviews

## A.4 Ethics From

**Central University Research Ethics Committee (CUREC)**

**CUREC 1A Form for Application form for research projects in the social sciences and humanities with less complex ethical issues**

The University of Oxford places a high value on the knowledge, expertise, and integrity of its members and their ability to conduct research to high standards of scholarship and ethics. The research ethics review process has been established to ensure that research involving human participants is conducted in a way that respects the dignity, rights, and welfare of participants, and minimises risk to participants, researchers, third parties, and to the University itself. It is assumed that all members of the University will take their responsibilities and obligations seriously, and will ensure that their research involving human participants is conducted according to established principles and good practice in their field and in accordance, where appropriate, with legal requirements.

Before completing this form, please refer to the guidance and flowchart on the Research Support website. Only type-written forms form will be accepted. Completed application forms should be emailed, along with relevant supporting documents, to your Departmental Research Ethics Committee (DREC) or to ethics@socsci.ox.ac.uk from your ox.ac.uk email address.

Please contact your DREC or the SSH IDREC if you have any questions about completing this form or the review process.

Advisory text is highlighted in yellow and should be deleted before finalising the document.

| **SECTION A: Filter for CUREC 2 application** | | |
|---|---|---|
| This section determines whether the application for ethics review should be made using the this form (CUREC 1A) or the CUREC 2 form (for research with more complex ethical issues). | | |
| **Please indicate with an 'X'.** | **Yes** | **No** |
| 1. Does the research involve the deception of participants? | ☐ | ☒ |
| 2. Are the research participants vulnerable in the context of the research, or classed as people whose ability to give free and informed consent is in question? For example,<br>• Participants aged 16 or under (also answer question A5);<br>• Participants aged 16 – 18 (refer to competent youths for guidance);<br>• adults at risk;<br>Note the University's Safeguarding Guidance and Code of Practice and its implications for researchers involving young people or adults at risk. | ☐ | ☒ |
| 3. By taking part in the research, will participants be at risk of criminal prosecution or significant harm? | ☐ | ☒ |
| 4. Does your research raise issues relevant to the Counter-Terrorism and Security Act (the Prevent Duty), which seeks to prevent people from being drawn into terrorism? Best Practice Guidance 07 on the Prevent Duty provides further guidance. | ☐ | ☒ |
| If you answered **'No'** to all the questions above, go to Section B. If you answered **'Yes'** to any question above, continue to question 5 below. | | |
| 5. Is your project covered by a CUREC Approved Procedure? | ☐ | ☐ |
| If yes, list the CUREC Approved Procedure(s) you will follow | | |
| If you answered '**Yes**' to ANY of questions 1-4, and answered '**No**' to question 5, **stop** completing this form and do not submit it for ethical review. You will instead need to submit a CUREC 2 application form. If you answered '**Yes**' to **any** of questions 1-4, and your project is covered by an | | |

| |
|---|
| Approved Procedure, **go on to Section B**. If more than one Approved Procedure applies, contact the SSH IDREC or your DREC for advice on whether a CUREC 2 form should be submitted instead. |

| **SECTION B: Researchers** | | |
|---|---|---|
| 1. Name of Principal Investigator or student's supervisor | Marina Jirotka | |
| 2. Department or Institute | Computer Science | |
| 3. University of Oxford telephone number | | |
| 4. University of Oxford email address | marina.jirotka@cs.ox.ac.uk | |
| Copy and paste the following six rows as necessary to complete for each additional researcher who will be involved in this study, including student(s) and those external to the University. | | |
| 5. Name of researcher or student | Towera Jessica Moyo | |
| 6. Department or Institute | Computer Science | |
| 7. University of Oxford telephone number | | |
| 8. University of Oxford email address | towera.moyo@new.ox.ac.uk | |
| 9. Role in research | Student | |
| 10. Degree programme, if student research | DPhil | |
| **The whole research team** | | |
| 11. Have the researchers undertaken research ethics and integrity training? | **Yes** ☒ | **No** ☐ |
| 12. Please provide details of any research ethics and integrity training undertaken, including the dates of the training. Alternatively state relevant research experience. | I have watched the pre-recorded research ethics video, provided by the link.<br><br>I have successfully completed the research integrity course on the provided link. | |
| 13. State any conflicts of interest and explain how these will be addressed. | | |

| **SECTION C: The research project** | |
|---|---|
| 1. Title of the research project | |
| Assessment of anti-trafficking software using a responsible innovation approach | |
| 2. Anticipated start date of the aspect of the research project involving human participants and/ or personal data (**dd/mm/yy**). | Date of ethics approval |
| 3. Anticipated research end date (**dd/mm/yy**). | June, 2022 |
| 4. Provide a brief lay summary of the aims and objectives of the research. This should cover the questions it will answer and any potential benefits. (max 300 words) | |

The aim of the project is to assess anti-trafficking software, identify the challenges associated with their use and development and the unintended consequences of using such software. Human trafficking is a global problem that affects people of all genders and ages. It is mainly committed in the real world but has now moved online due to the increase in the use of the internet and other technologies. Recently human traffickers have started using technology to recruit and exploit victims using websites, and social networking sites. As a response to this, different people such as technology experts, researchers, and anti-trafficking organizations are using technology to fight the malpractice. They are using different state of art technologies such as machine learning, data mining, and natural language processing to develop software tools that identify and track victims and traffickers online. Law enforcement and anti-trafficking organizations use these tools to aid their efforts of fighting human trafficking. However, there is currently no study that has reviewed the tools after they are deployed in the real world to identify challenges and unintended consequences of using them such as a high number of false positives, and potential bias in which certain groups of people are targeted. It is important to ensure that the tools take into account data ethics and informed consent, and do not breach any privacy rights or target certain groups of people. It is also necessary to review the development process of the tools to identify any challenges that may cause unfavourable results after the tools are deployed. The assessment of the tools will be guided by the AREA framework. It will highlight the challenges associated with their use and development, anticipate different kinds of unintended consequences, and identify strategies that can be put in place to mitigate the challenges and unintended consequences.

| 5. Please indicate the methods to be used (indicate with an '**X**'): | |
|---|---|
| Analysis of existing records | ☒ |
| Snowball sampling (recruiting through contacts of existing participants) | ☒ |
| Use of casual or local workers e.g. interpreters (refer to guidance in BPG 01: Researcher safety) | ☐ |
| Participant observation | ☐ |
| Covert observation | ☐ |
| Observation of specific organisational practices | ☐ |
| Participant completes questionnaire in hard copy | ☐ |

| | |
|---|---|
| Participant completes online questionnaire or other online task (refer to guidance in BPG 06: Internet-mediated research) | ☐ |
| Using social media to recruit or interact with participants (refer to guidance in BPG 06: Internet-mediated research) | ☐ |
| Participant performs paper and pencil task | ☐ |
| Participant performs verbal or aural task (e.g. for linguistic study) | ☐ |
| Focus group | ☐ |
| Interview (refer to guidance in BPG 10: Conducting research interviews) | ☒ |
| Audio recording of participant<br>(you will generally need specific consent from participants for this) | ☒ |
| Video recording of participant<br>(you will generally need specific consent from participants for this) | ☐ |
| Photography of participant<br>(you will generally need specific consent from participants for this) | ☐ |
| Others (please specify below) | ☐ |

6. Provide a brief summary of the research design and methods. What will research participants be asked to do? (max 300 words)
   Please also submit a copy of the questions participants will be asked, if applicable, or some information about the sorts of topics that will be covered.

I will interview developers of anti-trafficking software, and experts in anti-trafficking space.

The main topics to be covered in the interviews include:

- The challenges of developing and using anti-trafficking software
- Unintended consequences of identifying trafficking activity online and the strategies to minimize the consequences
- Recommendations on how to mitigate negative unintended consequences of anti-trafficking software

| | |
|---|---|
| 7. List the location(s) where the research will be conducted, including any other countries. | |
| 8. Clarify which parts of the research will be conducted in-person and which will take place remotely, e.g. online. | All the interviews will be conducted online. |

| | | |
|---|---|---|
| 9. If your research involves fieldwork or travel and your department requires a travel risk assessment, will you have completed and returned a risk assessment form beforehand? Please indicate with an '**X**'.<br>(This must be approved by your department before you travel. If you are travelling overseas, you are advised to take out University travel insurance.) | Yes | ☐ |
| | No | ☒ |
| | Not required | ☐ |

| | | |
|---|---|---|
| Refer to guidance available from your Department, the Safety Office, the Social Sciences Division, and the Humanities Division, and on travel for University business. | in this instance | |

| |
|---|
| 10. In the case of international or collaborative research, explain how you will address any ethical issues specific to the local context.<br>Please provide details of the local review, approval or permission obtained or required. Refer to the BPG 16: Social science research conducted outside the UK. If there will be no local review, explain why not.<br>Please also address any physical or psychological risks for Oxford researchers and local fieldworkers in Section G. |
| For collaborative research, the PI should keep a written record of evidence of whether ethics approval is also required by the collaborating institution(s) or whether the latter is prepared to rely on the ethics approval of the University of Oxford (where Oxford is the lead institution). Where ethics approval is required by the collaborating institution(s), copies of the other ethics applications (and confirmation of approval) should be submitted to the DREC or IDREC. |

| | |
|---|---|
| 11. Name of departmental/ peer reviewer (if applicable) | |
| 12. External organisation funding the research and grant reference (if applicable) | |
| 13. Please refer to the CUREC Best Practice Guidance and list any that have been used to develop your research. | |

## SECTION D: Recruitment of research participants

| | | |
|---|---|---|
| 1. Number of participants | 15 | |
| 2. How was the number of participants decided? | This has been done in order to get a range of opinions while supplementing the findings of additional methods used in the study. | |
| 3. Age range of participants | 24 and above | |
| 4. Inclusion criteria | Participants with substantive knowledge about human trafficking activity online, and individuals who have developed and used anti trafficking software. | |
| 5. Exclusion criteria | | |
| 6. Indicate with an '**X**' all intended recruitment methods<br>**Please submit copies of the recruitment material that will be used, e.g.** | Poster advert | ☐ |
| | Flyer | ☐ |
| | Email circulation | ☐ |
| | Social media (e.g. Twitter, Facebook) | ☐ |

| | | |
|---|---|---|
| **advertisement text, introductory email text.** | Website | ☒ |
| | In-person approach | ☐ |
| | Snowball sampling | ☒ |
| | Recruitment sites (e.g. Mechanical Turk) | ☐ |
| | Existing contacts or volunteer database | ☒ |
| | Other (please specify): | ☐ |
| 7. How will potential participants be identified and approached? | Majority of the participants will be known experts from our existing networks, and some will be identified through recommendation from other experts (snowball sampling). They will be mainly approached through email or LinkedIn. | |
| 8. Will informed consent be obtained from the research participants or their parents/ guardians? If not, please explain why not. | Yes – see supplemental information sheet and consent form. | |
| 9. For each activity or group of participants, explain how informed consent will be obtained from the participants themselves and/ or their parents/ guardians, if applicable.<br>How will their consent be recorded? | All the participants will be provided with an information sheet about the research, and a sample of the questions and topics that will be covered during the interview. This will help participants to decide whether they will participate in the study. If they choose to participate in the study, they will be given an online consent form to sign. The consent form will be shared via the GDPR-compliant email system.<br><br>Please submit copies of all participant-facing materials for review. E.g.:<br>• Recruitment material (e.g. emails, posters)<br>• Information for participants to read (or hear) before they agree to take part (e.g. written information or, if applicable, an outline oral information script).<br>• A document to record informed consent.<br>Further guidance and templates. | |
| 10. Provide details of any payments and incentives and the rationale for providing these. Further guidance in Best Practice | | |

| | |
|---|---|
| Guidance: 05 Payments and incentives in research. | |
| 11. Describe how participants<br>• may withdraw from the study<br>• may withdraw any personal information they have provided from the study<br>State any limits to withdrawal, for example once the data has been anonymised or at some other specified stage prior to publication. Make sure participants are aware of any withdrawal limits. | Participants may withdraw from the study by emailing the researcher (student) to state their wish to withdraw. While they may provide reasons for their withdrawal, this is not required. They can withdraw up to two weeks after data collection, and this will be communicated to them at the start of the study. |

## SECTION E: Research data

All information provided by participants is considered research data for the purpose of this form. Any research data from which participants can be identified is known as personal data; any personal data which is sensitive is considered special category data. Management of personal data, either directly or via a third party, must comply with the requirements of the UK General Data Protection Regulation (UK GDPR) and the Data Protection Act 2018, as set out in the University's Guidance on Data Protection and Research.

In answering the questions below, please also consider the points raised in the Data Protection Checklist and Data Protection Screening Assessment and whether, for higher-risk data processing, a separate Data Protection Impact Assessment may also be required for the research. Advice on research data management and security is available from Research Data Oxford and your local IT department. Advice on data protection is available from the Information Compliance team.

For guidance on conducting internet-mediated research, refer to CUREC's Best Practice Guidance 06: Internet-mediated research.

1. What data will be collected? (Indicate with an '**X**')

| | | | |
|---|---|---|---|
| Screening documents | ☐ | Task results (e.g. questionnaires, diaries) | ☐ |
| Consent records ( e.g., written consent forms, audio-recorded consent, assent forms) | ☒ | IP addresses (refer to Best Practice Guidance 09: Data collection, protection and management for guidance) | ☐ |
| Contact details for the purpose of this research only | ☒ | Field notes | ☐ |
| Contact details for future use (guidance) | ☒ | Photographs | ☐ |
| Opt-out forms | ☐ | Information about the health of the participant (including mental health) | ☐ |
| Audio recordings | ☒ | Previously collected (secondary) data | ☐ |
| Video recordings | ☐ | Data already in the public domain. Specify the source of the data: | ☐ |

| | | | |
|---|---|---|---|
| | | | |
| Transcript of audio/ video recordings | ☒ | Other, please specify: | ☐ |

| | |
|---|---|
| 2. During the course of the research, where will **each type of** research data be stored? | Give details of storage for each in terms of location and duration. List here each type of data selected above, and explain how each will be physically transferred (including movement of paper records and electronic downloads, such as for online surveys) from where it is collected to a suitable storage site (e.g. Nexus 365 OneDrive for Business, SharePoint, University servers). Do not store unencrypted data in freely available cloud services or unprotected USB drives Refer to the Best Practice Guidance on Internet-mediated research and Data collection, protection and management for guidance.<br><br>This study will have no paper records, and all the data will be collected online. Audio recording from the sessions will be done using a password-protected recorder, and any personal identifying information will be redacted when transcribing the data. Transcripts of data will also be stored on the university servers.<br><br>Consent forms, and interview data will be stored on the university servers. A spreadsheet that maps the participants to codes will be saved as a password protected document, stored on an external hard drive.<br><br>All anonymised data will be stored and accessed on a password-protected university machine. In addition, the data will be backed up using the university HFS. |
| 3. Who will have access to the research data during the project? | Give details of supervisors, colleagues and other researchers/ organisations (e.g., other universities, transcription services). Will individuals be identifiable from the data? Explain whether a unique participant number will be used instead of participant name. If identifiable data is to be shared with people external to the University of Oxford, explain how it will be transferred securely. Confidentiality agreements may be necessary if third parties are involved.<br><br>**Research team**<br>Professor Marina Jirotka (Supervisor)<br>Dr. Davina Durgana (Supervisor) |

| | | | |
|---|---|---|---|
| | Personal identifying information collected during the interviews will be removed, and participants will only be identified using numbered codes. | | |
| 4. Please complete this section if your research involves the use of secondary (i.e. previously collected) data. | **Please indicated with an 'X'.** | **Yes** | **No** |
| | Are data access agreements in place for access to and use of this secondary data? (If so, please attach these.) | ☐ | ☐ |
| | Did the individuals agree that their data could be used for this purpose? | ☐ | ☐ |
| | Could anyone (including members of the research team) link the data back to an individual or individuals? If this is a possibility, please explain how the associated ethical issues will be addressed: | ☐ | ☐ |
| 5. How do you intend to share the research data at the end of the project? | Depositing in a specialist data centre or archive | ☐ | |
| | Submitting to a journal to support a publication | ☐ | |
| | Depositing in an institutional repository | ☒ | |
| | Dissemination via a project or institutional website | ☐ | |
| | No plans to share the data | ☒ | |
| | Other (please specify): | ☐ | |
| 6. How do you intend to report and disseminate the results of the research? (Indicate with an 'X') | Thesis publication | ☒ | |
| | Publication in a peer reviewed journal | ☒ | |
| | Publicly available report | ☐ | |
| | Conference presentation | ☐ | |
| | Publication on a website | ☐ | |
| | Report to a research funder | ☐ | |
| | Providing participants with a lay summary of the results | ☐ | |
| | Submission for academic assessment | ☐ | |
| | Other (please specify): | ☐ | |
| 7. Explain what will happen to the data at the end of the research project. This question must be answered for each type of data, including completed consent forms. | | | |
| Please confirm that you will store research data safely for at **least 3 years** after final publication or public release and adhere to any additional research funder policies. Describe how and where the data will be stored, and if the researcher is likely to leave the University during this time, outline the plans for ensuring the data is handled in accordance with the detail provided in this form. | | | |

Explain any arrangements for making the data available for reuse, including any arrangements for archiving the data. Explain your approach to keeping any personal data. For example, if you wish to retain contact details in order to re-approach participants about future studies, you must explain this to them and obtain specific consent for this.
If it is no longer needed, or will not be archived, personal data should be destroyed in order to comply with the UK General Data Protection Regulation and the Data Protection Act, please confirm you will use a secure destruction process and confirm when and how each type of identifiable data will be destroyed.

All the data collected during the study will be stored securely for 3 years after the completion of the DPhil. The data will be stored on a secure deposit for the HCC research group.

| **SECTION F: Protection of research participants and their personal data** | | |
|---|---|---|
| 1. How identifiable will the participants be from the **research outputs**? (Indicate with an 'X') | Directly identifiable from the information included | ☐ |
| | Pseudonymised/ indirectly identifiable | ☐ |
| | Not identifiable – data is anonymous | ☒ |
| | Other, please specify: | ☐ |
| 2. To what extent will the **data** be de-identified? How identifiable will any individuals be from the research data? Describe any measures you will take towards assuring confidentiality, potential risks to confidentiality. | The information will be completely anonymised during the transcription process and will therefore not contain any information that can identify the participants. The participants will be identified using number codes. | |
| 3. How will you ensure that third parties (e.g., interpreters and transcribers) are aware of and adhere to the measures described in this form? | Confidentiality agreement templates are available via the Research Support website and guidance on the international transfer of personal data is available from the Information Compliance website.<br><br>N/A | |

## SECTION G: Risks and benefits of the research

1. Will the research involve topics that could be considered sensitive? If so:
    a. Please provide more detail or supporting information (such as the interview questions) to show the range of questions;
    b. Explain what steps will be taken to reduce risk of distress;
    c. Consider seeking advice from within your Department or from the ethics committee including whether the application might benefit from additional ethics review (e.g., via a CUREC 2 application).

The research does cover some sensitive topics. The following steps will be involved in reducing risk of participants

- Have several short breaks according to the mood of the participants when collecting data.
- Give them an option to continue the interview on another day.
- Notify them that they can withdraw from the study at any time during data collection and up to two weeks after completing data collection.

2. Describe any additional burden or risks to the participants and the steps you will take to address these.

Include other sorts of risk not mentioned elsewhere in your application, e.g., for projects in or near conflict zones.

3. Describe any physical or psychological risks to the researcher(s) (including local fieldworkers or research assistants) and the steps you will take to address these.

BPG 01: Researcher Safety contains guidance on addressing physical risks to researchers and the SSD website contains vicarious trauma resources.

- There will be regular check-ins with my supervisors, especially Dr Durgana as she has gone through a similar process while doing her PhD, and she is also currently working in this space.
- I will be journaling regularly during the process.
- I have also formed a small support group with a student from Harvard University who is currently working on a similar topic. We plan on having regular check-ups, and sharing tips on how to handle these types research. We plan to include more researchers working in this space.

4. Describe any benefits of the research, both to participants and to others.

The results will help technology experts to reflect on possible solutions to any potential challenges during the development of anti-trafficking software that target online sex strafficking. It will also

| |
|---|
| help the public and other interested parties to accept and have confidence in the tools that are developed. |
| 5. Give details of any other ethical issues or relevant information. |
| |

| **SECTION H: Professional guidelines** | | |
|---|---|---|
| Please indicate with an '**X**' at least one set of professional guidelines you will follow. | | |
| **Research specialism/ methodology** | **Association and guidance** | |
| Anthropology | Association of Social Anthropologists of the UK | ☐ |
| Computer Science | ACM Code of Ethics and Professional Conduct | ☐ |
| Criminology | British Society of Criminology Statement of Ethics | ☐ |
| Education | British Educational Research Association Ethical Guidelines for Educational Research | ☐ |
| Geography | American Association of Geographers Statement on Professional Ethics | ☐ |
| History | Oral History Society of the UK Ethical Guidelines | ☐ |
| Internet-mediated research | Association of Internet Researchers Ethical Guidelines<br>British Psychological Society: Ethics Guidelines for internet-mediated research<br>Association for Computing Machinery Code of Ethics and Professional Conduct | ☐ |
| Management | Academy of Management Code of Ethics | ☐ |
| Political Science | American Political Science Association (APSA) Guide to Professional Ethics in Political Science | ☐ |
| Politics | Political Studies Association. Guidelines for Good Professional Conduct | ☐ |
| Psychology | British Psychological Society Code of Ethics and Conduct | ☐ |
| Social research | Social Research Association: Ethical Guidelines | ☒ |
| Socio-legal studies | Socio-Legal Studies Association: Statement of Principles of Ethical Research Practice | ☐ |
| Sociology | The British Sociological Association: Statement of Ethical Practice | ☐ |
| Visual research | ESRC National Centre for Research Methods Review Paper: Visual Ethics: Ethical Issues in Visual Research | ☐ |
| Other professional guidelines | List any other guidelines used here. | ☐ |

## SECTION I: Endorsements and signatures

Please ensure this form is endorsed by the Principal Investigator (or student's supervisor), the Head of Department (or nominee) and, if student research, by the student themselves.
**The SSH IDREC Secretariat accepts either option below. If you have a DREC, check which signature option it prefers.**

- **Option 1: direct email endorsements**
  Each of the signatories should submit an email from a University of Oxford email address, indicating their acceptance of the responsibilities listed below.
- **Option 2: signatures**
  Please scan the signed form and email it to us as a PDF. Pasted images of signatures cannot be accepted.

**Endorsement by the Principal Investigator/ student supervisor and student, if applicable**

I/ we the researchers understand my/ our responsibilities as Principal Investigator (and student, if applicable) as outlined in the guidance on the CUREC website. I/ we declare that the answers above accurately describe the research as presently designed, and that the ethics committee will be informed of any changes to the project which affect the answers to this form.
I/ we will inform the relevant IDREC if the Principal Investigator changes.

| | |
|---|---|
| Name of Principal Investigator | |
| Principal Investigator's signature | Instead of a signature, endorsement may be provided by an email confirming the points above. |
| Date | |
| Name of student (if applicable) | |
| Student's signature | Instead of a signature, endorsement may be provided by an email confirming the points above. |
| Date | |

**Departmental endorsement** – from the Head of Department or nominee
(Another senior member of the department may sign where the head of department is the Principal Investigator, or where the Head of Department has appointed a nominee. Example nominees include Deputy Head of Department, Director of Research, or Director of Graduate/ Undergraduate Studies.)

I have read the research project application named above. On the basis of the information available to me, I:

- consider the PI and student researcher (if applicable) to be aware of their ethical responsibilities in regard to the ethical issues associated with this research;
- am satisfied that the proposed design and methodology are sound; the research has been subject to appropriate peer review and is likely to contribute to existing knowledge and/ or to the education and training of the researcher(s) and that it is in the public interest.

| | |
|---|---|
| Signature | Instead of a signature, endorsement may be provided by an email confirming the points above. |
| Name | |
| Role | |
| Date | |

# B

# Chapter 4 Supplementary Materials

## B.1 Data Cleaning Process

### B.1.1 Dataset 1

The following steps were involved in cleaning and preprocessing the data set:

1. Correcting spelling mistakes and spelling variations: The dataset exhibited numerous spelling errors and also had many spellings for the same word (e.g. 'Children' and 'Child-ren'). Corrections were made to correct these spelling mistakes and variations in columns such as profiles, destination, vulnerability_group_other_website, nationality, and industry.

2. Fixing Date and Time Columns: The date and time column, intercept_date, was formatted to ensure easy analysis. Additional columns were also generated to separate the data into day, month, and year components, simplifying subsequent analytical tasks.

3. Handling null data: Null cells within the dataset were addressed in two ways, depending on whether the cells contained numerical or non-numerical values. Data from numerical columns were filled with mean values, whereas data in non-numerical columns were filled with mode values.

4. Removing spaces and non-textual characters: Spaces and non-textual characters were systematically removed from the dataset to prevent potential errors during analysis.

5. Formatting Data: Data within the industry column were formatted to ensure a uniform casing, thereby facilitating grouping and analysis.

6. Grouping similar data: In the case of the destination column, some entries specified specific cities. For the analysis that focused on a national level, these entries were grouped under their respective country names.

These steps collectively ensured that the dataset was thoroughly cleaned and prepared for subsequent analysis.

### B.1.2 Dataset 2

The following steps were involved in the cleaning and preprocessing of the data set:

1. Removed columns not relevant to addressing the research questions in this sub-study. These typically involved legal actions, information about police stations, and police conduct.

2. Corrections were made to commonly found spelling errors and variations in columns, such as nationality and occupation. For example, "Ghananian" was changed to "Ghana" for nationality and "pupil" was modified to "student" for occupation.

3. Reformated the date and time columns (interview_date and intercept_date) into suitable date and time data types. In addition, new columns were created to separate the day, month, and year for better analysis.

4. Generated values for null cells. Similarly to dataset 1, the values were filled with the mean or the mode. Since most fields consisted of categorical textual data, the mode was used predominantly to address null cells.

5. Some columns were merged to improve analysis and easily identify patterns. These included columns related to recruitment methods, reasons for leaving, types of exploitation, and identification and travel information.

6. Removed spaces and non-textual characters to properly format text and avoid errors during analysis. Certain texts contained symbols or additional information in parentheses.

7. The extreme values were replaced by the mode of the respective field. For example, cases with extreme values, such as 5,000 traffickers for a single record, were substituted with the mode of the field.

8. Columns, such as nationality, were formatted to ensure a consistent case, allowing identical values in different cases to be represented by a single value.

# C

# Chapter 6 Supplementary Materials

## C.1 Annotation Guide

Thank you for your commitment to annotating the data for our human trafficking research project. Your role in labelling 1000 job posts is pivotal for creating a comprehensive dataset for academic purposes.

**Project Goal:**

The primary objective is to label 1000 job posts to establish a fully annotated dataset on potentially suspicious jobs related to human trafficking, contributing significantly to our academic research at the University of Oxford and future research projects.

**Instructions:**

**Step 1: Familiarize Yourself with Columns:**

- Job title
- City
- Job description
- Company Name
- Salary unit (hourly, monthly, yearly)
- Salary from (starting range)
- Salary to (ending range)
- Min Years of Experience (minimum required)
- Max Years (maximum required)
- Industry
- Status (you will be required to fill this field. It will have four options described below.)

**Step 2: Reviewing Jobs and fill in the 'status' column:**

- Assign one of the following labels after assessing the job based on your knowledge and investigative techniques:
    - **Not Suspicious:** Legitimate jobs with no concerns.
    - **Suspicious/Scam:** Ads showing less than 3 signs of illegitimacy.
    - **High Risk of Human Trafficking:** Significant signs of human trafficking.

**Considerations for Not Suspicious Label:**

- **Salary:** Compare with average; if it is normal range, then it causes no significant concern,  it aligns with normal specific job advertisements.
- **Experience:** If a job requires more than 3 years of experience, then it aligns with the job requirements.
- **Skills:** If a job emphasises relevant skills for a job, then it may prove legitimate.
- **Style of writing:** If the style is formal and in accordance with job advertisements, it aligns with normal job advertisements.

**Considerations for High-Risk Label, at least have more than 3 indicators available:**

- **Salary:** Compare with the average salary, determining if it is significantly higher or different may indicate potential trafficking.
- **Emphasis on No Experience:** Assess if the job seems unnecessary without relevant qualifications.
- **Industry:** Be cautious with industries associated with trafficking.
- **Compelling Package:** Watch for attractive offers like free transportation and food.
- **Style of Writing:** Look for grammatical errors and deceptive language.
- **Number of People Recruited:** Be wary of jobs recruiting more than ten people.

**Considerations for Suspicious Label, less than 3 indicators:**

The label is for job ads that are suspicious based on indicators such as salary but there is no sufficient evidence to be classied as high risk of trafficking. If the considerations for high risk are present but not more than 3 then it can indicate suspiciousness.

However, please note that these are considerations and you may use other indicators to help you classify a job ad. There are only meant to assist you during the process. Feel free to use other indicators not specified above, but please take note of them.

**Duration:**

The annotation period is set for three weeks.

**Confidentiality:**

## C.2 Chapter 6: Initial Data Collection Process

This section discusses the initial data collection process for the dataset. Data were collected from websites A and B, as previously mentioned in Chapter 6. It outlines the process of generating queries, collecting data, and preprocessing the data.

### C.2.1 Query Generator

Ten queries were generated based on the indicators identified in Chapters 4 and 5 to facilitate data collection. Five queries were executed on Website A, and identical queries were subsequently executed on Website B. All queries were performed with the location parameter set to Country A without specifying any particular city. The following queries were used to collect data from websites A and B.

1. Job advertisements seeking individuals without experience.
2. Job advertisements for cleaners.
3. Job advertisements for housekeeping positions.
4. Job advertisements for nannies.
5. Job targeting young people.

### C.2.2 Data Collection

Following the generation of the query list, a platform called Browse AI (BrowseAI, n.d.) was used for data extraction, monitoring, and scheduling of data extraction based on specified time periods. Browse AI was selected because of its time efficiency. The primary components extracted from the websites included date (GMT), job link, search results (URL), job title, company name, location, salary, job description, posted date and job link. The data were stored in an Excel spreadsheet after extraction.

The objective was to label 1,000 job advertisements, comprising 500 from Website A and 500 from Website B. A threshold of 100 advertisements was established

for each query. For queries that yielded more than 100 results, randomisation was implemented to select 100 job advertisements. Certain queries, such as those pertaining to nannies and youth, retrieved only a limited number of posts. Consequently, a script was developed to execute the queries daily for a week on both websites to ensure the acquisition of sufficient data.

### C.2.3 Data Preprocessing

Prior to the labelling process, the data were cleaned to ensure compatibility with the format required for human annotators. Different processes were implemented to clean the data acquired from Websites A and B owing to the variation in formats between the two sources.

For Website A website, the data were cleaned by :

1. The short and full descriptions contained excessive double or triple spacing after extraction. This was corrected to ensure single spacing.

2. Some job advertisements included additional text in the 'location' field, which was removed.

For Website B, the data were preprocessed by:

1. The location and company name were extracted together in a single field due to the design of the website, so it was necessary to separate them and drop the company name.

2. Standardising the format of the company name by applying title case to ensure consistency across all roles.

Ethics approval was denied during the preprocessing stage, leading to the use of a third-party dataset, as described in Section 6.6.4. All data records from this process were deleted.

# References


International Labour Organisation, Walk Free, & IOM UN Migration. (2022). *Global Estimates of Modern Slavery Forced Labour and Forced Marriage* (tech. rep.).

United Nations Office on Drugs and Crime. (2023). *GLOBAL REPORT ON TRAFFICKING IN PERSONS 2022.* UNITED NATIONS.

Miller, A. D. (2014). HUMAN SEX TRAFFICKING : INDIVIDUAL RISK FACTORS FOR RECRUITMENT , TRAFFICKING , AND VICTIMIZATION ON THE INTERNET A Thesis Presented to the Faculty of California State University Dominguez Hills In Partial Fulfillment of the Requirements for the Degree Ma.

Dorton, K. (2019). THE EVOLUTION OF HUMAN TRAFFICKING: THE USE OF SOCIAL MEDIA AND FINANCIAL INSTITUTIONS. (December).

OSCE. (2020). *Leveraging innovation to fight trafficking in human beings : A comprehensive analysis of technology tools.*

Satyarthi, K. (2019, July). OPINION: Child sex trafficking in the high-speed internet era. https://news.trust.org/item/20190725142903-y1h0m/

Latonero, M., & Movius, L. (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). http://technologyandtrafficking.usc.edu/files/2011/09/HumanTrafficking_FINAL.pdf

Department of State, U. (2023). *OFFICE TO MONITOR AND COMBAT T RAFFICKING IN PERSONS Online Recruitment of Vulnerable Populations for Forced Labor* (tech. rep.). https://www.state.

Progetto Tenda, Danish Refugee Council, Nesta Italia, LABC Srl, & CWEP. (2021). Free2Link Report.

Anthony, B. (2018). On-Ramps, Intersections, Routes: and Exit A Roadmap for Systems and Industries to Prevent and Disrupt Human Trafficking.

Feehs, K., & Wheeler, A. C. (2020). Federal Human Trafficking Report.

Ramchandani, P., Bastani, H., & Wyatt, E. (2021). *Unmasking human trafficking risk in commercial sex supply chains with machine learning* (tech. rep.). https://ssrn.com/abstract=3866259

United Nations Office on Drugs and Crime. (2020). Global Report on Trafficking in Persons.

Calderwood, I. (2018, September). Human Traffickers Using Facebook to Lure Migrants Into 'Trips'. https://www.globalcitizen.org/en/content/human-trafficking-facebook-adverts-refugees/

Sykiotou, A. (2007). Trafficking in human beings: Internet recruitment. *Europe*.

United Nations Office on Drugs and Crime. (2018). *Global Report on Trafficking in Persons* (tech. rep.).

Europol. (2014). Intelligence Notfification: Trafficking in Human Beings and the Internet. (October), 1–3. https://www.europol.europa.eu/publications-documents/trafficking-in-human-beings-and-internet

Lockett, J. (2018, July). Inside Philippines' £1bn 'made to order' child sex abuse industry making sick vids for Brit paedos. https://www.thesun.co.uk/news/6868807/philippines-child-sex-abuse-vids-brit-paedos/

Mitchell, K. J., & Boyd, D. (2014). Understanding the role of technology in the commercial sexual exploitation of children: the perspective of law enforcement. *Crimes Against Children Research Center*. http://scholars.unh.edu/cgi/viewcontent.cgi?article=1036&context=ccrc

Hultgren, M., Whitney, J., Jennex, M. E., & Elkins, A. (2018). A knowledge management approach to identify victims of human sex trafficking. *Communications of the Association for Information Systems*, *42*(1), 602–620.

Dubrawski, A., Miller, K., Barnes, M., Boecking, B., & Kennedy, E. (2015). Leveraging Publicly Available Data to Discern Patterns of Human-Trafficking Activity. *Journal of Human Trafficking*, *1*(1), 65–85.

Mayorga, M., & Özaltın, O. (2023). *Margaret Tobey Dissertation.tex* (tech. rep.).

Lugo-Graulich, K., Meyer, L. F., Souza, K., Tapp, S. N., Maryfield, B., & Bostwick, L. (2024). Improving Sex Trafficking Victim Identification: Indicators of Trafficking in Online Escort Ads. *Journal of Human Trafficking*.

Alvari, H., Shakarian, P., & Snyder, J. E. K. (2017). Semi-supervised learning for detecting human trafficking. *Security Informatics*, *6*(1).

Giommoni, L., & Ikwu, R. (2024). Identifying human trafficking indicators in the UK online sex market. *Trends in Organized Crime*, *27*(1), 10–33.

Spotlight. (n.d.). Spotlight: Human Trafficking Intelligence and Leads | Thorn. https://www.thorn.org/spotlight/

Marinus Analytics. (2022). Traffic Jam — Marinus Analytics. https://www.marinusanalytics.com/traffic-jam

Flores, I. (2022, August). Google AI Mistakenly Flags Parent for Child Sexual Abuse Material After Taking Photos of Toddler for Medical Use | Tech Times. https://www.techtimes.com/articles/279466/20220821/google-ai-flags-parent-child-sexual-abuse-material-medical-use.htm

Centric. (2022). EPOOLICE.

Thorn. (2018). Survivor Insights:The Role of Technology in Domestic Minor Sex Trafficking JANUARY. (Thorn).

Greiman, V., & Bain, C. (2013). The Emergence of Cyber Activity. *International Journal of Cyber Warfare and Terrorism ( IJCWT )*, *12*(2), 41–49.

Ibanez, M., & Suthers, D. D. (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*, 1556–1565.

Latonero, M. (2012). The Rise of Mobile and the Diffusion of Technology-Facilitated Trafficking. *SSRN Electronic Journal*, (November).

Zhou, S., Peng, J., & Ferrara, E. (2024). *Tracing the Unseen: Uncovering Human Trafficking Patterns in Job Listings* (tech. rep.). https://www.chineseinla.com/user/id

Volodko, A., Cockbain, E., & Kleinberg, B. (2020). " Spotting the signs " of trafficking recruitment online : exploring the characteristics of advertisements targeted at migrant job-seekers.

Musto, J., Thakor, M., & Gerasimov, B. (2020). Editorial: Between hope and hype: Critical evaluations of technology's role in anti-trafficking. *Anti-Trafficking Review*, *14*, 1–14.

Leary, M. G. (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*.

Hultgren, M., Jennex, M. E., Persano, J., & Ornatowski, C. (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), 4344–4353.

Whitney, J. (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.

Hernandez-Alvarez, M. (2019). Detection of possible human trafficking in twitter. *Proceedings - 2019 International Conference on Information Systems and Software Technologies, ICI2ST 2019*, 187–191.

Simonson, E. L. (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.

Sarkar, S. (2015). Use of technology in human trafficking networks and sexual exploitation: A cross-sectional multi-country study. *Transnational Social Review*, *5*(1), 55–68.

Deeb-Swihart, J., Endert, A., & Bruckman, A. (2019). Understanding law enforcement strategies and needs for combating human trafficking. *Conference on Human Factors in Computing Systems - Proceedings*.

Arenas, M., Corcho, O., Simperl, E., Strohmaier, M., D'Aquin, M., Srinivas, K., Groth, P., Dumontier, M., Heflin, J., Thirunarayan, K., & Staab, S. (2015). The semantic Web -ISWC 2015: 14th international semantic web conference bethlehem, 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*, 205–221.

Whitney, J., Jennex, M., Elkins, A., & Frost, E. (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*, 4273–4282.

Tong, E., Jones, C., Zadeh, A., & Morency, L. P. (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*, 1547–1556.

Wang, H., Cai, C., Philpot, A., Latonero, M., Hovy, E. H., & Metzler, D. (2012). Data integration from open internet sources to combat sex trafficking of minors. *ACM International Conference Proceeding Series*, 246–252.

Silva, D. R., Philpot, A., Sundararajan, A., Bryan, N. M., & Hovy, E. (2014). Data integration from open internet sources and network detection to combat underage sex trafficking. *ACM International Conference Proceeding Series*, 86–90.

Alvari, H., Shakarian, P., & Snyder, J. E. (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*(Tvpa), 133–138.

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

Wang, Laber, E., Saanchi, Y., & Caltagirone, S. (2020). Sex trafficking detection with ordinal regression neural networks. *arXiv*, (April 2018).

Paula, A., Terron, A., Salazar, J. Y., Rivas, P., Caballero, E. Q., & Rodriguez Perez, A. (2023). *Task-Specific or Task-Agnostic? A Statistical Inquiry into BERT for Human Trafficking Risk Prediction* (tech. rep.). https://huggingface.co/sbcBI/sentiment_analysis_model

Esfahani, S. S., Cafarella, M. J., Pouyan, M. B., DeAngelo, G. J., Eneva, E., & Fano, A. E. (2020). Context-specific language modeling for human trafficking detection from online advertisements. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference]*, 1180–1184.

Liu, J., Yu, H., Sujaya, V., Nair, P., Pelrine, K., Rabbany, R., & Link, G. (2023). *SWEET-Weakly Supervised Person Name Extraction for Fighting Human Trafficking Example* (tech. rep.).

Summers, L., Shallenberger, A. N., Cruz, J., & Fulton, L. V. (2023). A Multi-Input Machine Learning Approach to Classifying Sex Trafficking from Online Escort Advertisements. *Machine Learning and Knowledge Extraction*, *5*(2), 460–472.

Wiriyakun, C., & Kurutach, W. (2021). Feature Selection for Human Trafficking Detection Models. *Proceedings - 2021 IEEE/ACIS 21st International Fall Conference on Computer and Information Science, ICIS 2021-Fall*, 131–135.

Vajiac, C., Olligschlaeger, A., Li, Y., Nair, P., Lee, M.-C., Park, N., Rabbany, R., Horng Chau, D., & Faloutsos, C. (2022). *TRAFFICVIS: Fighting Human Trafficking through Visualization* (tech. rep.).

Vajiac, C., Lee, M. C., Kulshrestha, A., Levy, S., Park, N., Olligschlaeger, A., Jones, C., Rabbany, R., & Faloutsos, C. (2023). DeltaShield: Information Theory for Human- Trafficking Detection. *ACM Transactions on Knowledge Discovery from Data*, *17*(2).

Nair, P., Li, Y., Vajiac, C., Olligschlaeger, A., Lee, M. C., Park, N., Chau, D. H., Faloutsos, C., & Rabbany, R. (2022). VisPaD: Visualization and Pattern Discovery for Fighting Human Trafficking. *WWW 2022 - Companion Proceedings of the Web Conference 2022*, 273–277.

Ardakani, H. M. (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*.

Shorten, A., & Smith, J. (2017). Mixed methods research: Expanding the evidence base. *Evidence-Based Nursing*, *20*(3), 74–75.

Tashakkori, A., & Creswell, J. W. (2007). Editorial: The New Era of Mixed Methods.

Yu, X., & Khazanchi, D. (2017). Using embedded mixed methods in studying is phenomena risks and practical remedies with an illustration. https://unomaha.az1.qualtrics.com/jfe/form/

Creswell J. W. & Plano Clark, V. L. (2006). *CHOOSING A MIXED METHODS DESIGN* (tech. rep.).

O'Cathain, A., Murphy, E., & Nicholl, J. (2010). Three techniques for integrating data in mixed methods studies. *BMJ*, *341*(7783), 1147–1150.

Campana, P. (2022). *Online and technology-facilitated trafficking in human beings Summary and recommendations Group of Experts on Action against Trafficking in Human Beings* (tech. rep.).

Justice and Home Affairs Committee. (2022). *Technology rules? The advent of new technologies in the justice system* (tech. rep.). http://www.parliament.uk/hlregister

Jick, T. D. (1979). *Mixing Qualitative and Quantitative Methods: Triangulation in Action* (tech. rep. No. 4).

European Commission. (2020). Responsible research & innovation | Horizon 2020. https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation

UK Research and Innovation. (2023). Research integrity in healthcare technologies. https://www.ukri.org/councils/epsrc/guidance-for-applicants/what-to-include-in-your-proposal/health-technologies-impact-and-translation-toolkit/research-integrity-in-healthcare-technologies/responsible-research-and-innovation/

Owen, R., Stilgoe, J., Macnaghten, P., Gorman, M., Fisher, E., & Guston, D. (2013). A Framework for Responsible Innovation. *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, (April), 27–50.

Jirotka, M., Barbara, G., Stahl, B., Eden, G., & Hartswood, M. (2017). Innovation in the digital age. *China's Quest for Innovation*, 195–215.

Franco-Santos, M., & Otley, D. (2018). Reviewing and Theorizing the Unintended Consequences of Performance Management Systems. *International Journal of Management Reviews*, *20*, 696–730.

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

Allais, C. (2006). Overview of trafficking.

U.S. Congress. (2000). *Victims of Trafficking and Violence Protection Act of 2000* (tech. rep.).

Crawford, M. (2017). International Sex Trafficking. *Women and Therapy*, *40*(1-2), 101–122.

International Labour Office. Special Action Programme to Combat Forced Labour. (2008). *Combating forced labour : a handbook for employers & business*. ILO.

United Nations: Office on Drugs and Crime. (2023). *GLOBAL REPORT ON TRAFFICKING IN PERSONS 2022.* UNITED NATIONS.

Arronte, R. (2018). A partner in TRAFFICKING: THE ROLE OF INTERNET TECHNOLOGIES IN THE FACILITATION OF HUMAN TRAFFICKING. (December).

United Nations Office on Drugs and Crime. (2012). *Global Report on Trafficking in Persons*. http://www.unodc.org/unodc/en/data-and-analysis/glotip.html

US Department of State. (2019). TRAFFICKING IN PERSONS REPORT.

UNICEF. (2003). *Trafficking in human beings, especially women and children, in Africa.*

Davis, H. (2013). Defining 'Pimp': Working Towards a Definition in Social Research. *Sociological Research Online*, 18 (1) 11.

Mitchell, K. J., Jones, L. M., Finkelhor, D., & Wolak, J. (2011). Internet-facilitated commercial sexual exploitation of children: Findings from a nationally representative sample of law enforcement agencies in the United States. *Sexual Abuse: Journal of Research and Treatment*, *23*(1), 43–71.

Riley, B. (2016, July). The '5 disguises' used by human traffickers - CNN. https://edition.cnn.com/2016/07/13/opinions/5-disguises-human-trafficker/index.html

Fonrouge, G. (2018, April). The sick tactics sex traffickers use to find victims. https://nypost.com/2018/04/17/how-sex-traffickers-hunt-for-victims-and-brainwash-them/

Konrad, R. A., Trapp, A. C., Palmbach, T. M., & Blom, J. S. (2016). Overcoming human trafficking via operations research and analytics: Opportunities for methods, models, and applications. *European Journal of Operational Research*, *259*(2), 733–745.

Sykiotou, A. (2017). Cyber trafficking: recruiting victims of human trafficking through the net. http://www.state.gov/g/drl/rls/hrrpt/2010/eap/154382

US Department of State. (2020). Trafficking in Persons Report.

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/

Olivette. (2021). Human Trafficking | Olivette, MO. https://www.olivettemo.com/426/Human-Trafficking

Gerry QC, F., Muraszkiewicz, J., & Vavoula, N. (2016). The role of technology in the fight against human trafficking: Reflections on privacy and data protection concerns. *Computer Law and Security Review*, *32*(2), 205–217.

Rodriguez, C. (2018, November). Craigslist and Backpage: The Websites Used for Sex Trafficking – StMU History Media. https://stmuhistorymedia.org/craigslist-and-backpage-sex-trafficking-at-your-fingertips/

Siddiqui, S. (2016). *Social Media its Impact with Positive and Negative Aspects* (tech. rep. No. 2). www.ijcat.com

Arli, D. (2017). Does Social Media Matter? Investigating the Effect of Social Media Features on Consumer Attitudes. *Journal of Promotion Management*, *23*(4), 521–539.

Petrosyan, S. (2022). What Is a Job Board? (Definition and Examples). https://www.jobboardfire.com/academy/what-is-a-job-board

Walby, S., Apitzsch, B., Armstrong, J. E., Balderston, S., Szmagalska-Follis, K., L Francis, K., B. J., May-Chahal, C. A., Rashid, A., Shire, K., Towers, J., & Tunte, M. (2016). Study on the gender dimension of trafficking in human beings, 18–31.

Allen, C. (2019, March). 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/

Shepherd, J. (2024). 33 Essential Facebook Statistics You Need To Know In 2024. https://thesocialshepherd.com/blog/facebook-statistics#:~:text=There%20are%203.065%20Billion%20Monthly%20Active%20Users%20on%20Facebook&text=This%20is%20an%20increase%20from,they%20had%202.96%20billion%20users.

Europol. (2020). The challenges of countering human trafficking in the digital era, 1–7.

Freedom United. (2018, October). Sex Trafficking Survivor Sues Facebook - FreedomUnited.org. https://www.freedomunited.org/news/sex-trafficking-survivor-sues-facebook/

Xiao, C., Freeman, D. M., & Hwa, T. (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*, 91–102.

Blomberg, M. (2019, April). Global taskforce tackles cybersex child trafficking in the Philippines | Reuters. https://www.reuters.com/article/us-philippines-trafficking-children/global-taskforce-tackles-cybersex-child-trafficking-in-the-philippines-idUSKCN1RR1D1

Li, R., Tobey, M., Mayorga, M. E., Caltagirone, S., & Özaltın, O. Y. (2023). Detecting Human Trafficking: Automated Classification of Online Customer Reviews of Massage Businesses. *Manufacturing and Service Operations Management*, *25*(3), 1051–1065.

Pendergrass, M. A. (2018). IS ACTIVE CYBER DEFENSE THE ANSWER ? by William Lloyd A Capstone Project Submitted to the Faculty of Utica College May 2018 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Cybersecurity. (May).

TellFinder. (n.d.). Home | TellFinder Alliance. https://www.tellfinderalliance.com/

Brewster, B., Polovina, S., Rankin, G., & Andrews, S. (2014). Knowledge management and human trafficking: Using conceptual knowledge representation, text analytics and open-source data to combat organized crime. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8577 LNAI*(8577), 104–117.

idTraffickers. (2022). Software for law enforcement that identifies enslaved persons and traffickers. https://www.idtraffickers.com/

NORC. (n.d.). Sex Trafficking Operations Portal (STOP) | NORC.org. https://www.norc.org/Research/Projects/Pages/sex-trafficking-operations-portal-stop.aspx

Freedom Signal. (n.d.). Freedom Signal. https://freedomsignal.org/

TraffickCam. (2022). About TraffickCam. https://traffickcam.com/about

Musto, J. (2020). *The Limits and Possibilities of Data-Driven Antitrafficking Efforts* (tech. rep.). https://readingroom.law.gsu.edu/gsulrhttps://readingroom.law.gsu.edu/gsulr/vol36/iss4/11

Ediae, C., & Kuteesa, P. N. (2024). PREDICTIVE ANALYTICS FOR PROACTIVE SUPPORT IN TRAFFICKING PREVENTION AND VICTIM REINTEGRATION. *Engineering Science & Technology Journal*, *5*(4), 1502–1523.

Rivas, P., Cerny, T., Rodriguez Perez, A., Turek, J. S., Giddens, L., Bichler, G., & Petter, S. (2024). *On the Challenges of Creating Datasets for Analyzing Commercial Sex Advertisements to Assess Human Trafficking Risk and Organized Activity* (tech. rep.).

Anderson, P., Barker, K., Curtis, J., Davidson, J., Ellis, D. A., Fiani, C., Ghai, S., Khamis, M., Livingstone, S., Martellozzo, E., Mcgill, M., Orben, A., Rashid, A.,

& Tantam, W. (2024). *Ofcom Consultation: Protecting children from harms online* (tech. rep.).

Deeb-Swihart, J., Endert, A., & Bruckman, A. (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), 1–29.

Taylor, R. L., & Latonero, M. (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. http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf

Amin, S. (2010). A Step Towards Modeling and Destabilizing Human Trafficking, 2–7.

Freeman, N. K., Keskin, B. B., & Bott, G. J. (2022). Collaborating with Local and Federal Law Enforcement for Disrupting Sex Trafficking Networks. *INFORMS Journal on Applied Analytics*, *52*(5), 446–459.

Mucinno, E. (2019, June). Active Learning for Fast Data Set Labeling | by Eric Muccino | Mindboard | Medium. https://medium.com/mindboard/active-learning-for-fast-data-set-labeling-890d4080d750

Jacobs, P. F., Wenniger, G. M. d. B., Wiering, M., & Schomaker, L. (2021). Active learning for reducing labeling effort in text classification tasks. (550), 1–24. http://arxiv.org/abs/2109.04847

TELUS International. (2021, June). What Is Data Annotation And Why Does It Matter? https://www.telusinternational.com/articles/what-is-data-annotation

Arora, V. (2019). Supervised, Unsupervised and Semi-supervised ML. https://rstudio-pubs-static.s3.amazonaws.com/559023_f62f3bce1be14fb9b248127194c0c1e3.html

Dickson, B. (2021, January). What is semi-supervised machine learning? – TechTalks. https://bdtechtalks.com/2021/01/04/semi-supervised-machine-learning/

Bruijn, L. (2020, July). 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

Widmann, M. (2020, September). Cohen's Kappa: what it is, when to use it, how to avoid pitfalls | KNIME. https://www.knime.com/blog/cohens-kappa-an-overview

Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., & Leuven, K. U. (2010). *A METHOD BASED ON TEMPORAL CONCEPT ANALYSIS FOR DETECTNG AND PROFILING HUMAN TRAFFICKING SUSPECTS* (tech. rep.).

Barber, J. (n.d.). Latent Dirichlet Allocation (LDA) with Python. https://rstudio-pubs-static.s3.amazonaws.com/79360_850b2a69980c4488b1db95987a24867a.html

Madhavan, S. (2020, January). Deep learning architectures – IBM Developer. https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2015). Gated feedback recurrent neural networks. *32nd International Conference on Machine Learning, ICML 2015*, *3*, 2067–2075.

Burbano, D., & Hernandez-Alvarez, M. (2018). Identifying human trafficking patterns online. *2017 IEEE 2nd Ecuador Technical Chapters Meeting, ETCM 2017*, *2017-Janua*, 1–6.

Diaz, M., & Panangadan, A. (2020a). Natural Language-based Integration of Online Review Datasets for Identification of Sex Trafficking Businesses. *Proceedings - 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science, IRI 2020*, 259–264.

Diaz, M., & Panangadan, A. (2020b). Natural Language-based Integration of Online Review Datasets for Identification of Sex Trafficking Businesses. *Proceedings - 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science, IRI 2020*, 259–264.

Cotterill, D. (2023). The Impact of Social Media, the Internet, and Legislation on Online Minor Sex Trafficking. *Journal of Digital Life and Learning*, *3*(2), 18–45.

Flynn, M. (2021). FOSTA-SESTA and its impact on sex workers - AIDS United. https://aidsunited.org/fosta-sesta-and-its-impact-on-sex-workers/

Romano, A. (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

GOV.UK. (2023). A guide to the Online Safety Bill - GOV.UK. https://www.gov.uk/guidance/a-guide-to-the-online-safety-bill

ActiveFence. (2024). Compliance and Regulation. https://www.activefence.com/compliance-regulation/

OneTrust Data Guidance. (2022). Australia: The scope of the Online Safety Act | Insights | DataGuidance. https://www.dataguidance.com/opinion/australia-scope-online-safety-act

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.

Gazette, G., & Notice, G. (2022). *Films and Publications Act: Regulations: Amendment* (tech. rep.). www.gpwonline.co.za

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

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.

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

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/

Gezinski, L. B., & Gonzalez-Pons, K. M. (2022). Sex Trafficking and Technology: A Systematic Review of Recruitment and Exploitation. *Journal of Human Trafficking*.

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2024, May). A comprehensive survey on applications of transformers for deep learning tasks.

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need* (tech. rep.).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. http://arxiv.org/abs/1810.04805

ORBIT. (2021). The Keys of RRI | Orbit RRI. https://www.orbit-rri.org/resources/keys-of-rri

Pili. (2017). Responsible Research Innovaton. http://2017.igem.org/Team:Exeter/HP/Silver

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580.

Yaghmaei, E., & van de Poel, I. R. (2020). *Assessment of Responsible Innovation.*

Owen, R. (2014). The UK Engineering and Physical Sciences Research Council's commitment to a framework for responsible innovation. *Journal of Responsible Innovation*, *1*(1), 113–117.

Etikan Ilker, Musa, S., & Alkassim Rukayya. (2016). Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1.

Naderifar, M., Goli, H., & Ghaljaie, F. (2017). Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research. *Strides in Development of Medical Education*, *14*(3).

Barclay, C. (2018). *Semi-Structured Interviews* (tech. rep.). http://sociology.fas.harvard.edu/files/sociology/files/interview_strategies.pdf

Sampson Helen. (2004). Navigating the waves: the usefulness of a pilot in qualitative research.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Clarke, V., & Braun, V. (2017). Thematic analysis. *Journal of Positive Psychology*, *12*(3), 297–298.

Braun, V., Clarke, V., Hayfield, N., & Terry, G. (2019). Thematic analysis BT - Handbook of research methods in health social sciences. *Springer Nature Singapore Pte Ltd.*, 843–860. https://doi.org/10.1007/978-981-10-5251-4_103

Lauri, M. A. (2011). *Triangulation of Data Analysis Techniques* (tech. rep.). http://www.psych.lse.ac.uk/psr/%5D

Jonsen, K., & Jehn, K. A. (2009). Using triangulation to validate themes in qualitative studies. *Qualitative Research in Organizations and Management: An International Journal*, *4*(2), 123–150.

Home Office. (2018, July). Modern Slavery Act 2015 - GOV.UK. https://www.gov.uk/government/collections/modern-slavery-bill#guidance-for-enforcement-officers

Deeb-Swihart, J. (2022). DEEB-SWIHART-DISSERTATION-2022.

Wright, G. (n.d.). What is Access Control? https://www.techtarget.com/searchsecurity/definition/access-control

Deepak, G. (2023). 3.1.1 Problem Definition. Full Series… | by Goel Deepak | Medium. https://medium.com/@goel_deepakin/mastering-the-art-of-problem-definition-in-machine-learning-and-ai-a838f8837ce2

Kniazieva, Y. (2022). Data Collection as A Major Asset in the Machine Learning Pipeline | Label Your Data. https://labelyourdata.com/articles/data-collection-methods-AI

O'Regan, S. (n.d.). AI Problem Statements — Simon O'Regan. https://www.simonoregan.com/short-thoughts/ai-problem-statements

Mumuni, A., & Mumuni, F. (2022, December). Data augmentation: A comprehensive survey of modern approaches.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021, November). Data and its (dis)contents: A survey of dataset development and use in machine learning research.

Hube, C., Fetahu, B., & Gadiraju, U. (2019). Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. *Conference on Human Factors in Computing Systems - Proceedings*.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., & Oliva, A. (2020). Moments in Time Dataset: One Million Videos for Event Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(2), 502–508.

He, H., Zha, S., & Wang, H. (2019). Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual. http://arxiv.org/abs/1908.10763

Teney, D., Abbasnedjad, E., & Hengel, A. v. d. (2020). Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision. http://arxiv.org/abs/2004.09034

Seneor, A., & Mezzanotte, M. (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/

McKenna, M. (n.d.). Bias in AI: How to Mitigate Bias in AI Systems | Toptal®. https://www.toptal.com/artificial-intelligence/mitigating-ai-bias

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. http://arxiv.org/abs/1810.01943

Khani, F., & Liang, P. (2020). Removing Spurious Features can Hurt Accuracy and Affect Groups Disproportionately. http://arxiv.org/abs/2012.04104

Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M., & Farkash, A. (2022). Data minimization for GDPR compliance in machine learning models. *AI and Ethics*, *2*(3), 477–491.

DataGrail. (2023, February). How to Implement Data Minimization | DataGrail. https://www.datagrail.io/blog/data-privacy/how-to-implement-data-minimization/

Al-Zubaidie, M., Zhang, Z., & Zhang, J. (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).

Langarizadeh, M., Orooji, A., & Sheikhtaheri, A. (2018). Effectiveness of anonymization methods in preserving patients' privacy: A systematic literature review. *Studies in Health Technology and Informatics*, *248*, 80–87.

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. (2016). Towards the Science of Security and Privacy in Machine Learning. http://arxiv.org/abs/1611.03814

Anthonysamy, P., Rashid, A., Chitchyan, R., & Lancaster, S. (2017). *Privacy Requirements: Present & Future* (tech. rep.). https://www.whatsapp.com/faq/general/26000016

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., & Choi, Y. (2020). Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. http://arxiv.org/abs/2009.10795

Wang, A. (2020). *REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets* (tech. rep.). https://github.com/princetonvisualai/revise-tool.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., & Choi, Y. (2019). *WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale* (tech. rep.). http://winogrande.allenai.org.

Egnyte. (2021). Data Auditing – Improve Data Quality | Egnyte. https://www.egnyte.com/guides/governance/data-auditing

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., Van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., & Sokolov, A. (2022). Quality at a Glance: An Audit ofWeb-Crawled Multilingual Datasets. *Bonaventure F. P. Dossou*, *12*, 47.

Annah Whyton, M. (2014). *Template for information sheet* (tech. rep.).

Burkart, N., & Huber, M. F. (2021). *A Survey on the Explainability of Supervised Machine Learning* (tech. rep.).

Freitas, A. A. (2014). *Comprehensible Classification Models-a position paper* (tech. rep.).

Maveli, N. (n.d.). Spectra - Demystifying Post-hoc Explainability for ML models. https://spectra.mathpix.com/article/2021.09.00007/demystify-post-hoc-explainability

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (n.d.). *A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME* (tech. rep.).

Vimbi, V., Shaffi, N., & Mahmud, M. (2024, December). Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024, January). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence.

Bai, T., Luo, J., Zhao, J., Wen, B., & Wang, Q. (2021). Recent Advances in Adversarial Training for Adversarial Robustness. http://arxiv.org/abs/2102.01356

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021, January). A Comprehensive Survey on Transfer Learning.

Donges, N. (2019, June). The Random Forest Algorithm: A Complete Guide | Built In. https://builtin.com/data-science/random-forest-algorithm

Soni, D. (2022, September). Feedback Loops in Machine Learning Systems | by Devin Soni | Towards Data Science. https://towardsdatascience.com/feedback-loops-in-machine-learning-systems-701296c91787

C3.AI. (n.d.). Feedback Loop. https://c3.ai/glossary/features/feedback-loop/

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022, October). A survey of human-in-the-loop for machine learning.

Metaxa, D., & Hancock, J. (2022). *KEY TAKEAWAYS* (tech. rep.).

Digital Regulation Cooperation Forum. (2022, September). Auditing algorithms: the existing landscape, role of regulators and future outlook - GOV.UK.

https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054.

Xin, D., Ma, L., Liu, J., Macke, S., Song, S., & Parameswaran, A. (2018). Learning efficiently over heterogeneous databases: Sampling and constraints to the rescue. *Proceedings of the 2nd Workshop on Data Management for End-To-End Machine Learning, DEEM 2018 - In conjunction with the 2018 ACM SIGMOD/PODS Conference*.

UXPIN. (n.d.). What is UX Honeycomb and How it Impacts User Experience? https://www.uxpin.com/studio/blog/ux-honeycomb-definition-and-use/

Von Schomberg, R. (2013). *A Vision of Responsible Research and Innovation*.

Galal, S. (2024). Africa: total population 2000-2030 | Statista. https://www.statista.com/statistics/1224168/total-population-of-africa/

Boye, G. (2022, February). Africa - World Music Guide - LibGuides at Appalachian State University. https://guides.library.appstate.edu/world-music/africa

Obokata, T. (2019, May). Human Trafficking in Africa. In *The african court of justice and human and peoples' rights in context* (pp. 529–552). Cambridge University Press.

Fabbri, C., Stöckl, H., Zimmerman, C., Jones, K., Cook, H., Galez-Davis, C., Grant, N., & Lo, Y. (2023). Labor Recruitment and Human Trafficking: Analysis of a Global Trafficking Survivor Database.

Britton, H. E., & Dean, L. A. (2014). Policy Responses to Human Trafficking in Southern Africa: Domesticating International Norms. *Human Rights Review*, *15*(3), 305–328.

Children International. (2023). Poverty in Africa.

Macrotrends. (2025). Sub-Saharan Africa Unemployment Rate 1991-2023 | MacroTrends. https://www.macrotrends.net/countries/SSF/sub-saharan-africa-/unemployment-rate

Onuoha, B. (2011). The state human trafficking and human rights issues in Africa. *Contemporary Justice Review: Issues in Criminal, Social, and Restorative Justice*, *14*(2), 149–166.

UN Trade & Development. (2021). Facts and Figures | UNCTAD. https://unctad.org/press-material/facts-and-figures-7#:~:text=While%20in%202019%2C%20478%20million,was%20projected%20without%20the%20pandemic.

UNESCO. (n.d.). Education in Africa | UNESCO UIS. https://uis.unesco.org/en/topic/education-africa

Fedele, A. (2021, January). Human Trafficking Trends in Sub-Saharan Africa (Infographic) | ASEC-SLDI News. http://asec-sldi.org/news/current/human-trafficking-sub-saharan-africa/

Global Slavery Index. (2018). Africa | Region Highlights. https://www.globalslaveryindex.org/2018/findings/regional-analysis/africa/

Aparicio, M., Org, M. A., & Costa, C. J. (2014). *Data Visualization* (tech. rep.).

Great Learning. (2024). Understanding Data Visualization Techniques. https://www.mygreatlearning.com/blog/understanding-data-visualization-

techniques/#:~:text=Data%20visualization%20is%20a%20graphical,outliers%2C%20and%20patterns%20in%20data.

Hashemi-Pour, C. (2024). What is data visualization and why is it important? https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization#:~:text=Data%20visualization%20is%20the%20practice,outliers%20in%20large%20data%20sets.

Eberhard, K. (2023). The effects of visualization on judgment and decision-making: a systematic literature review. *Management Review Quarterly*, *73*(1), 167–214.

Yada, V. (2022, December). How Can Python Be Used for Data Visualization? - KDnuggets. https://www.kdnuggets.com/2022/12/python-used-data-visualization.html#:~:text=Python%20has%20easy%20syntaxes%20and,are%20used%20for%20data%20visualization.

Pandas. (2024). pandas - Python Data Analysis Library. https://pandas.pydata.org/

Panel. (2024). Overview — Panel v1.4.5. https://panel.holoviz.org/

HoloViews. (2024). Installation — HoloViews v1.19.1. https://holoviews.org/

Matplotlib. (2024). Matplotlib — Visualization with Python. https://matplotlib.org/

GeoPandas. (2024). Plotting with Folium — GeoPandas 1.0.1+0.g747d66e.dirty documentation. https://geopandas.org/en/stable/gallery/plotting_with_folium.html

Statista Research Department. (2024). Human trafficking victims globally 2022 | Statista. https://www.statista.com/statistics/459637/number-of-victims-identified-related-to-labor-trafficking-worldwide/

The World Bank. (2022, October). African Governments Urgently Need to Restore Macro-Economic Stability and Protect the Poor in a Context of Slow Growth, High Inflation. https://www.worldbank.org/en/news/press-release/2022/10/04/african-governments-urgently-need-to-restore-macro-economic-stability-and-protect-the-poor-in-a-context-of-slow-growth

Nayupe, S. F., Touray, A., Tembo, D., Msiska, M. T., Mputeni, J., & Muhome, A. H. (2023). Addressing the rights and well-being of domestic workers in Africa: Status in Malawi. *Public Health Challenges*, *2*(3).

Vaidya, D. (2023). Chi Square Test - Meaning, Formula, Examples, Independence. https://www.wallstreetmojo.com/chi-square-test/

McLeod, S. (n.d.). Understanding P-Values and Statistical Significance. https://www.simplypsychology.org/p-value.html

Foster, G., Lane, D., Scott, D., Hebl, M., Guerra, R., Osherson, D., & Zimmer, H. (2024). 11.8: Post Hoc Tests - Statistics LibreTexts. https://stats.libretexts.org/Bookshelves/Applied_Statistics/An_Introduction_to_Psychological_Statistics_(Foster_et_al.)/11%3A_Analysis_of_Variance/11.08%3A_Post_Hoc_Tests

Sharpe, D. (2015). Chi-Square Test is Statistically Significant: Now What? *Practical Assessment, Research, and Evaluation Practical Assessment*, *20*.

Durso, C. (n.d.). Contingency Tables in Statistics | Definition & Examples - Lesson | Study.com. https://study.com/academy/lesson/contingency-table-statistics-probability-examples.html#:~:text=A%20contingency%20table%20is%20a%20data%20table%20that%20compares%20two,the%20two%20sets%20are%20correlated.

Frost, J. (n.d.). Contingency Table: Definition, Examples & Interpreting - Statistics By Jim. https://statisticsbyjim.com/basics/contingency-table/

Pandas. (2025). pandas.crosstab — pandas 2.2.3 documentation. https://pandas.pydata.org/docs/reference/api/pandas.crosstab.html

Schnell, A. (n.d.). Chi-Square test: Do you need 5 observations per cell? https://www.theanalysisfactor.com/chi-square-test-of-independence-rule-of-thumb/

Bobbitt, Z. (2021). The Four Assumptions of a Chi-Square Test. https://www.statology.org/chi-square-test-assumptions/

JMP, S. D. (2023). Chi-Square Test of Independence. https://www.jmp.com/en_be/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html

Kavuma, R. (2011, October). Free universal secondary education in Uganda has yielded mixed results | Global development | The Guardian. https://www.theguardian.com/global-development/poverty-matters/2011/oct/25/free-secondary-education-uganda-mixed-results

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*, *20*(3), 973–989.

Starfish Malawi. (n.d.). Education in Malawi. https://starfishmalawi.com/education-in-malawi

Cross Catholic Outreach. (2021). Problems With Education in Malawi and How to Solve Them. https://crosscatholic.org/blogs/2021/12/problems-with-education-in-malawi-and-how-to-solve-them/

Botha, B. (n.d.). New Pathways to Commercialize Agriculture in Malawi. https://blogs.worldbank.org/en/nasikiliza/new-pathways-commercialize-agriculture-malawi#:~:text=The%20World%20Bank's%202022%20Poverty, even%20higher%2C%20at%2060%25.

Desmon, S. (2022, September). CCP Finds Success in Reducing Child Marriage in Malawi. https://ccp.jhu.edu/2022/09/19/married-child-stopping-malawi/

Gumbi, K. (2023, August). South Africa's unemployment rate drops marginally in second quarter | Reuters. https://www.reuters.com/world/africa/south-africas-unemployment-rate-down-slightly-326-q2-2023-08-15/#:~:text=PRETORIA%2C%20Aug%2015%20(Reuters),2021%2C%20data%20showed%20on%20Tuesday.

Trading Economics. (2023a). South Africa Youth Unemployment Rate - 2023 Data - 2024 Forecast. https://tradingeconomics.com/south-africa/youth-unemployment-rate

Trading Economics. (2023b). South Africa Unemployment Rate - 2023 Data - 2024 Forecast - 2000-2022 Historical. https://tradingeconomics.com/south-africa/unemployment-rate

Embassy of the replublic of Kenya. (n.d.). The education system in Kenya. https://kenemb.ru/en/page/education-system.html

Britannica. (2024). Kenya - Education, Literacy, Schools | Britannica. https://www.britannica.com/place/Kenya/Education

UNICEF. (2023). Education | UNICEF Kenya. https://www.unicef.org/kenya/education

Nzamba Kitonga Advocates LLP. (n.d.). EMPLOYMENT RIGHTS: DOMESTIC WORKERS. https://kitllp.com/employment-rights-domestic-workers/

Macro Trends. (2023). Angola Literacy Rate 2001-2023. https://www.macrotrends.net/countries/AGO/angola/literacy-rate

Horowitz Research. (2023). Horowitz Study Reveals Perceptions of Ad Effectiveness by Age; Among Younger Americans, Digital Platforms More Effective Than Ads on

Live and Streamed TV - Horowitz Research. https://www.horowitzresearch.com/all/horowitz-study-reveals-perceptions-of-ad-effectiveness-by-age/

Shawky, S., Kubacki, K., Dietrich, T., & Weaven, S. (2019). Using social media to create engagement: a social marketing review. *Journal of Social Marketing*, *9*(2), 204–224.

World Health Organisation. (2023). Youth-led campaigns: fresh ways of communicating about public health issues. https://www.who.int/azerbaijan/news/item/11-08-2023-youth-led-campaigns-fresh-ways-of-communicating-about-public-health-issues

Tyler, H. (2022). Article: The Increasing Use of Artificial Intellig.. | migrationpolicy.org. https://www.migrationpolicy.org/article/artificial-intelligence-border-zones-privacy

Williams, K. (2022, December). Participant Observation 101: Definition, Types, Uses, Examples. https://surveysparrow.com/blog/participant-observation/#:~:text=Passive%20participant%20observation,done%20without%20the%20participants'%20knowledge.

Murthy, D. (2008). Digital ethnography: An examination of the use of new technologies for social research.

Thomas, D. R. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, *27*(2), 237–246.

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, *2*, 8–14.

Erlingsson, C., & Brysiewicz, P. (2017, September). A hands-on guide to doing content analysis.

Spicker, P. (2011). Ethical covert research. *Sociology*, *45*(1), 118–133.

Thompson, A., Stringfellow, L., Maclean, M., & Nazzal, A. (2020). *A NOTE ON VERSIONS Ethical considerations and challenges for using digital ethnography to research vulnerable populations* (tech. rep.). http://hdl.handle.net/10871/120310

Marzano, M. (2021, December). Covert Research Ethics.

Chae, Y., & Davidson, T. (2023). *LARGE LANGUAGE MODELS FOR TEXT CLASSIFICATION: FROM ZERO-SHOT LEARNING TO FINE-TUNING Large language models for text classification* (tech. rep.).

Lee, M.-c., Levy, S., Park, N., Jones, C., & Faloutsos, C. (2021). INFOSHIELD : Generalizable Information-Theoretic Human-Trafficking Detection.

Caers, R., De Feyter, T., De Couck, M., Stough, T., Vigna, C., & Du Bois, C. (2013, September). Facebook: A literature review.

ActiveFence. (2023). Advancing Trust Safety.

Perry, C. ( r., Chhatralia, K., Damesick, D., Hobden, S., Volpe, L., & Health Foundation (Great Britain). (2015). *Behavioural insights in health care : nudging to reduce inefficiency and waste*.

Twitter. (n.d.). Introduction. https://communitynotes.x.com/guide/en/about/introduction

Khanna, V. (2024). Choose Your AI Weapon: Deep Learning or Machine Learning. https://shelf.io/blog/choose-your-ai-weapon-deep-learning-or-traditional-machine-learning/#:~:text=Traditional%20machine%20learning%20models%20generally,smaller%20datasets%20or%20simpler%20algorithms.

Explainable AI. (2024). Deep Learning vs. Traditional Machine Learning: Explore the differences between deep learning and traditional ML techniques, including use cases. https://www.linkedin.com/pulse/deep-learning-vs-traditional-machine-explore-differences-fxcwc/

Garrido-Merchan, E. C., Gozalo-Brizuela, R., & Gonzalez-Carvajal, S. (2023). Comparing BERT against Traditional Machine Learning Models in Text Classification. *Journal of Computational and Cognitive Engineering*.

Kaggle. (n.d.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/

Parti, A. (2024). What is Inter-Rater Reliability? (Examples and Calculations). https://pareto.ai/blog/inter-rater-reliability#:~:text=High%20inter%2Drater%20reliability%20in,and%20more%20effective%20AI%20performance.

Fendler, W. P., Barrio, M., Spick, C., Allen-Auerbach, M., Ambrosini, V., Benz, M., Bluemel, C., Grewal, R. K., Lapa, C., Miederer, M., Nicolas, G., Schuster, T., Czernin, J., & Herrmann, K. (2017). 68Ga-DOTATATE PET/CT interobserver agreement for neuroendocrine tumor assessment: Results of a prospective study on 50 patients. *Journal of Nuclear Medicine*, *58*(2), 307–311.

Burchfiel, A. (2022). What is NLP (Natural Language Processing) Tokenization? - tokenex. https://www.tokenex.com/blog/ab-what-is-nlp-natural-language-processing-tokenization/

Ashraf, A. (2024). Tokenization in NLP : All you need to know | by Abdallah Ashraf | Medium. https://medium.com/@abdallahashraf90x/tokenization-in-nlp-all-you-need-to-know-45c00cfa2df7#:~:text=Tokenization%20helps%20disambiguate%20the%20text,accuracy%20of%20subsequent%20NLP%20tasks.

NLTK. (n.d.). NLTK :: nltk.tokenize package. https://www.nltk.org/api/nltk.tokenize.html

Tum, P. (2020, February). A Survey of the State-of-the-Art Language Models up to Early 2020 | by Phylypo Tum | Medium. https://medium.com/@phylypo/a-survey-of-the-state-of-the-art-language-models-up-to-early-2020-aba824302c6

Brownlee, J. (2019, August). What Are Word Embeddings for Text? https://machinelearningmastery.com/what-are-word-embeddings/

Hoque, M. (2023). A Comprehensive Overview of Transformer-Based Models: Encoders, Decoders, and More | by Minhajul Hoque | Medium. https://medium.com/@minh.hoque/a-comprehensive-overview-of-transformer-based-models-encoders-decoders-and-more-e9bc0644a4e5

Suri, M. (2022). A Dummy's Guide to Word2Vec. Essentials of Word2Vec + Implementing... | by Manan Suri | Medium. https://medium.com/@manansuri/a-dummys-guide-to-word2vec-456444f3c673

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. http://arxiv.org/abs/1301.3781

Rong, X. (2014). [1411.2738] word2vec Parameter Learning Explained. https://ar5iv.labs.arxiv.org/html/1411.2738

Subedi, N. (2018, July). FastText: Under the Hood. Where we look at how one of the best... | by Nishan Subedi | Towards Data Science. https://towardsdatascience.com/fasttext-under-the-hood-11efc57b2b3

Ganesan, K. (n.d.). FastText vs. Word2vec: A Quick Comparison - Kavita Ganesan, Ph.D. https://kavita-ganesan.com/fasttext-vs-word2vec/#.YZVnUtDP02w

Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation* (tech. rep.). http://nlp.

Gupta, V., Saw, A. K., Talukdar, P. P., & Netrapalli, P. (2019). Unsupervised Document Representation using Partition Word-Vectors Averaging | OpenReview. https://openreview.net/forum?id=HyNbtiR9YX

Kujlzb, J. (2024). GPT 4 vs GPT 3: Top Differences That You Should Know in 2024. https://www.geeksforgeeks.org/gpt-4-vs-gpt-3/

Sruthi, R. (2024). Understand Random Forest Algorithm (Updated 2024). https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Breiman, L. (2001). *Random Forests* (tech. rep.).

Tabsharani, F. (2023). What is a support vector machine? | Definition from WhatIs. https://www.techtarget.com/whatis/definition/support-vector-machine-SVM

Brownlee, J. (2021). One-vs-Rest and One-vs-One for Multi-Class Classification - MachineLearningMastery.com. https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/

Wadhwatanya. (2021). Multi-class Classification — One-vs-All & One-vs-One | by Wadhwatanya | Medium. https://wadhwatanya1234.medium.com/multi-class-classification-one-vs-all-one-vs-one-993dd23ae7ca

Dobson, J. A., & Barnett, G. A. (2018). *An Introduction to Generalized Linear Models Fourth Edition* (tech. rep.).

Saini, A. (2024). Random Forest Algorithm for Beginners in Data Science | Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/

Scikit Learn. (n.d.). LogisticRegression — scikit-learn 1.5.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

M, J. (2020). Understanding Regularization in Logistic Regression | by Jun M. | Towards Data Science. https://towardsdatascience.com/understanding-regularization-in-machine-learning-5a0369ac73b9

RITHP. (2023). Logistic Regression and regularization: Avoiding overfitting and improving generalization | by RITHP | Medium. https://medium.com/@rithpansanga/logistic-regression-and-regularization-avoiding-overfitting-and-improving-generalization-e9afdcddd09d

Piccialli, V., & Sciandrone, M. (2022). Nonlinear optimization and support vector machines. *Annals of Operations Research*, *314*(1), 15–47.

Chauhan, V. K., Dahiya, K., & Sharma, A. (2019, August). Problem formulations and solvers in linear SVM: a review.

DeAngelo, G., Shapiro, J. N., Borowitz, J., Cafarella, M., Ré, C., & Shiffman, G. (2019). Pricing risk in prostitution: Evidence from online sex ads. *Journal of Risk and Uncertainty*, *59*(3), 281–305.

Yiu, T. (2019, June). Understanding Random Forest. How the Algorithm Works and Why it Is... | by Tony Yiu | Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Koehrsen, W. (2018, August). An Implementation and Explanation of the Random Forest in Python | by Will Koehrsen | Towards Data Science. https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76

Brownlee, J. (2023). A Gentle Introduction to k-fold Cross-Validation - MachineLearningMastery.com. https://machinelearningmastery.com/k-fold-cross-validation/

Lyashenko, V., & Jha, A. (2024). Cross-Validation in Machine Learning: How to Do It Right. https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right

Nagaraj, Y. (2023). Stratified K-Fold Cross-Validation: An In-depth Look. https://www.linkedin.com/pulse/stratified-k-fold-cross-validation-in-depth-look-yeshwanth-n/

Bajaj, A. (2023). Performance Metrics in Machine Learning [Complete Guide] - neptune.ai. https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide

Jain, A. (2024). A Comprehensive Guide to Performance Metrics in Machine Learning | by Abhishek Jain | Medium. https://medium.com/@abhishekjainindore24/a-comprehensive-guide-to-performance-metrics-in-machine-learning-4ae5bd8208ce

Gupta, L. (2020, November). Differences Between Word2Vec and BERT | by Lavanya Gupta | The Startup | Medium. https://medium.com/swlh/differences-between-word2vec-and-bert-c08a3326b5d1

Bhakta, S. (2024). How to Generate Word Embedding using BERT? - GeeksforGeeks. https://www.geeksforgeeks.org/how-to-generate-word-embedding-using-bert/

Joy, S., Dofadar, D., Khan, R., Ahmed, S., & Rahman, R. (2022). [2208.01355] A Comparative Study on COVID-19 Fake News Detection Using Different Transformer Based Models. https://ar5iv.labs.arxiv.org/html/2208.01355

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. http://arxiv.org/abs/1909.11942

Kacprzak, K. (2024). RoBERTa vs. BERT: Exploring the Evolution of Transformer Models - Data Engineering, MlOps and Databricks services. https://dsstream.com/roberta-vs-bert-exploring-the-evolution-of-transformer-models/#:~:text=BERT%20has%20been%20used%20to,detection%20in%20social%20media%20texts.

Jorge, L. (2023). GPT-3, BERT, and RoBERTa | AI Model Analysis & Comparison | by liva jorge | Medium. https://medium.com/@livajorge7/gpt-3-bert-and-roberta-ai-model-analysis-comparison-7dfab049367d

Khanna, C. (2021). Byte-Pair Encoding: Subword-based tokenization | Towards Data Science. https://towardsdatascience.com/byte-pair-encoding-subword-based-tokenization-algorithm-77828a70bee0

Sanbit. (2024). Byte-Pair Encoding (BPE) in NLP - GeeksforGeeks. https://www.geeksforgeeks.org/byte-pair-encoding-bpe-in-nlp/

Adith, N. T. (n.d.). Embeddings in BERT. https://iq.opengenus.org/embeddings-in-bert/

Raman, K. (2023). Introduction to BERT and Segment Embeddings. https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-bert/

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. http://arxiv.org/abs/1907.11692

Huang, Z., Liang, D., Xu, P., & Xiang, B. (2020). Improve Transformer Models with Better Relative Position Embeddings. http://arxiv.org/abs/2009.13658

Efimov, V. (2023). Large Language Models: RoBERTa — A Robustly Optimized BERT Approach | by Vyacheslav Efimov | Towards Data Science. https://towardsdatascience.com/roberta-1ef07226c8d8

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher Richard, Amatriain, X., & Gao, J. (2023). [2402.06196] Large Language Models: A Survey. https://ar5iv.labs.arxiv.org/html/2402.06196v2

Kalyan, S. K. (2023). [2310.12321] A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4. https://ar5iv.labs.arxiv.org/html/2310.12321

Kublik, S., & Saboo, S. (2022). A Beginner's Guide to GPT-3 | DataCamp. https://www.datacamp.com/blog/a-beginners-guide-to-gpt-3

Kelly, W. (2024). GPT-3.5 vs. GPT-4: Biggest differences to consider | TechTarget. https://www.techtarget.com/searchenterpriseai/tip/GPT-35-vs-GPT-4-Biggest-differences-to-consider

Bobbitt Zach. (2021). What is Inter-rater Reliability? (Definition & Example). https://www.statology.org/inter-rater-reliability/

Japkowicz, N., & Stephen, S. (2002). *The class imbalance problem: A systematic study 1* (tech. rep.). IOS Press.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Abhinav, R. (2023). Improving Class Imbalance with Class Weights in Machine Learning | by Ravi Abhinav | Medium. https://medium.com/@ravi.abhinav4/improving-class-imbalance-with-class-weights-in-machine-learning-af072fdd4aa4#:~:text=Calculating%20class%20weights%20is%20an,majority%20class%20during%20model%20training.

Sobhanam, H., & Prakash, J. (2023). Analysis of fine tuning the hyper parameters in RoBERTa model using genetic algorithm for text classification. *International Journal of Information Technology (Singapore)*, *15*(7), 3669–3677.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. http://arxiv.org/abs/2001.08361

Wang, Z., Pang, Y., & Lin, Y. (2023). Large Language Models Are Zero-Shot Text Classifiers. http://arxiv.org/abs/2312.01044

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. http://arxiv.org/abs/2005.14165

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., & Wang, G. (2023). Text Classification via Large Language Models. http://arxiv.org/abs/2305.08377

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2023). Open-Source LLMs for Text Annotation: A Practical Guide for Model Setting and Fine-Tuning. http://arxiv.org/abs/2307.02179

Springbord. (2023). Challenges Of Data Labelling And How To Overcome Them – Springbord. https://www.springbord.com/blog/challenges-of-data-labeling/

SuperAI. (2023). What is Domain Knowledge? https://www.getsuper.ai/post/what-is-domain-knowledge

Shen, L., Sun, Y., Yu, Z., Ding, L., Tian, X., & Tao, D. (2023). On Efficient Training of Large-Scale Deep Learning Models: A Literature Review. https://ar5iv.labs.arxiv.org/html/2304.03589

Malhotra, S., & Godayal, D. (2018, June). An introduction to part-of-speech tagging and the Hidden Markov Model. https://www.freecodecamp.org/news/an-introduction-to-part-of-speech-tagging-and-the-hidden-markov-model-953d45338f24/

Toloka. (2023). Revolutionizing Data Labeling with Large Language Models: The Future of Data Annotation. https://toloka.ai/blog/automated-data-labeling-with-llms/

Kotwel. (n.d.). Ensuring Labeling Quality in Machine Learning. https://kotwel.com/ensuring-labeling-quality-in-machine-learning/

Pičuljan, N. (n.d.). Advanced Data Labeling Methods for Machine Learning | Toptal®. https://www.toptal.com/machine-learning/advanced-data-labeling-methods

Salma, S., & Liu, X. (2019). Overfitting Mechanism and Avoidance in Deep Neural Networks. https://ar5iv.labs.arxiv.org/html/1901.06566

Bejani, M. M., & Ghatee, M. (2021). A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review*, *54*(8), 6391–6438.

Roy, S., Mehera, R., Pal, R. K., & Bandyopadhyay, S. K. (2023). Hyperparameter optimization for deep neural network models: a comprehensive study on methods and techniques. *Innovations in Systems and Software Engineering*.

Qi, X., & Xu, B. (2023). Hyperparameter optimization of neural networks based on Q-learning. *Signal, Image and Video Processing*, *17*(4), 1669–1676.

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). *THE COMPUTATIONAL LIMITS OF DEEP LEARNING* (tech. rep.).

Talaei Khoei, T., Ould Slimane, H., & Kaabouch, N. (2023, November). Deep learning: systematic review, models, challenges, and research directions.

Chen, C., Dell, N., & Roesner, F. (2019). Computer security and privacy in the interactions between victim service providers and human trafficking survivors. *Proceedings of the 28th USENIX Security Symposium*, 89–104.

Chamberlain, L. (2019). *FOSTA: A Hostile Law with a Human Cost* (tech. rep.).

Simon, C. (2018). Perma | On Backpage | Tits and Sass. https://perma.cc/39AT-6MKT

Marshall, R. (2016). *Sex Workers and Human Rights, A Critical Analysis of Laws Sex Work* (tech. rep.). http://perma.cc/A8BDZNUR%5D.

Erin Taylor. (2019, December). Sex Workers Are at the Forefront of the Fight Against Mass Surveillance and Big Tech.

Tichenor, E. (2020). 'I've Never Been So Exploited': The consequences of FOSTA-SESTA in Aotearoa New Zealand. *Anti-Trafficking Review*, *2020*(14), 99–115.

Peersman, C., Edwards, M., Williams, E., & Rashid, A. (2022). Automatic User Profiling in Darknet Markets: a Scalability Study. http://arxiv.org/abs/2203.13179

Peersman, C. (2018). Detecting Deceptive Behaviour in the Wild :

Rodriguez, A. (n.d.). [2311.13118] Combatting Human Trafficking in the Cyberspace: A Natural Language Processing-Based Methodology to Analyze the Language in Online Advertisements. https://ar5iv.labs.arxiv.org/html/2311.13118v1

Bonnet, A. (2023). Inter-rater Reliability: Definition & Applications | Encord. https://encord.com/blog/inter-rater-reliability/

Park, S., Wang, A. Y., Kawas, B., Liao, Q. V., Piorkowski, D., & Danilevsky, M. (2021). Facilitating Knowledge Sharing from Domain Experts to Data Scientists for Building NLP Models. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 585–596.

Unterkalmsteiner, M., & Yates, A. (2023). Expert-sourcing Domain-specific Knowledge: The Case of Synonym Validation. https://ar5iv.labs.arxiv.org/html/2309.16798

Alsmadi, I., & Gan, K. H. (2019, June). Review of short-text classification.

Roldós, I. (n.d.). Major Challenges of Natural Language Processing (NLP). https://monkeylearn.com/blog/natural-language-processing-challenges/

Waters, P. (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

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation.

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

Iguazio. (n.d.). What is Concept Drift. https://www.iguazio.com/glossary/concept-drift/

Ryan, C., Zeplowitz, D., Julie Boatright Wilson, P., & Goodman, J. (2016). *Partnerships in Investigating Sex Trafficking Bridging Gaps to Support Survivors Prepared for Polaris Faculty Advisors* (tech. rep.).

Trafficking Analysis Hub. (n.d.). Academic Institutions | Traffik Analysis Hub. https://www.traffikanalysis.org/academic-insitutions

BrowseAI. (n.d.). Scrape and Monitor Data from Any Website with No Code. https://www.browse.ai/