CS6220 Project Proposal: Understanding Victim-Perpetrator Dynamics and Patterns of Exploitation in Human Trafficking

Leo Senai, Sihui Lin

Abstract(to be updated later)

1. Introduction

Human trafficking remains one of the most severe global issues, affecting millions of men, women, and children every year across both developing and developed countries. Victims vary in age, race, gender, and nationality, but they are often from marginalized or vulnerable communities.

Our project analyzes the patterns and dynamics of human trafficking by utilizing synthetic data from the Counter-Trafficking Data Collaborative (CTDC). By integrating these datasets, we seek to uncover how relationships between victims and traffickers (e.g., family members, intimate partners, or strangers) influence the types of exploitation and control mechanisms deployed. Specifically, we will investigate which demographic profiles are most vulnerable to forced labor versus sexual exploitation, and how recruitment roles and control mechanisms vary across these profiles. Additionally, expanding our geographic scope allows us to conduct comparative analyses across regions (e.g., U.S. vs. Asia) and evaluate the effects of trafficking laws before and after their implementation.

We aim to understand the complex relationships between victims and perpetrators, as well as the varied methods of recruitment and control used in these exploitative acts. With the insights gained from the results, we can make recommendations for policy decisions and intervention strategies, provide better support for victims, contribute to more effective prevention efforts, and combat human trafficking on a global scale.

2. Problem Description

The dynamics of exploitation vary significantly across regions, types of exploitation, and victim demographics. Analyzing these dynamics is helpful for understanding which factors contribute to different types of exploitation, identifying the most vulnerable victim profiles, and determining how relationships between victims and traffickers impact the methods of control used. This research aims to explore the following questions:

- How do trends in different types of exploitation (e.g., forced labor vs. sexual exploitation) evolve over time and across regions?

- What is the relationship between victim demographics (such as age, gender) and specific types of exploitation?
- How do patterns of exploitation differ geographically?

3. Data Sources

- Summary of data source: CTDC victim-perpetrator synthetic data (https://www.ctdatacollaborative.org/global-victim-perpetrator-synthetic-dataset)
    - High level summary: data from over 17,000 victims and survivors of trafficking from across 123 countries/territories.
    - Also includes accounts of 37,000 perpetrators.
- One or two sentences about synthetic data and how it was generated
    - "This dataset is generated with the [differential privacy algorithm](https://github.com/microsoft/synthetic-data-showcase) developed at Microsoft Research."

This study utilizes the Victim-Perpetrator Synthetic Dataset from the Counter-Trafficking Data Collaborative (CTDC). The dataset includes detailed information on over 17,000 victims and survivors of human trafficking from 123 countries/territories and records on approximately 37,000 perpetrators. This dataset represents one of the most comprehensive compilations of trafficking data available for analysis.

The data is synthetically generated using differential privacy algorithms developed at Microsoft Research. This process ensures the privacy of individuals by maintaining statistical properties of the data while anonymizing specific details. This enables researchers to analyze patterns and relationships in trafficking without exposing sensitive information.

The dataset includes:

- Victim demographic information such as gender, age, and region of origin.
- Perpetrator details including roles, relationships with victims, and control mechanisms.
- Types of exploitation experienced by victims, such as forced labor or sexual exploitation.
- Geographic and temporal patterns of trafficking activities.

For this project, we focus specifically on three key regions of exploitation: Europe, Asia, and Africa. This allows for comparative analyses across these diverse regions while maintaining a manageable scope for analysis and prediction tasks.

4. Approach to Solution

Data Cleaning and Preparation

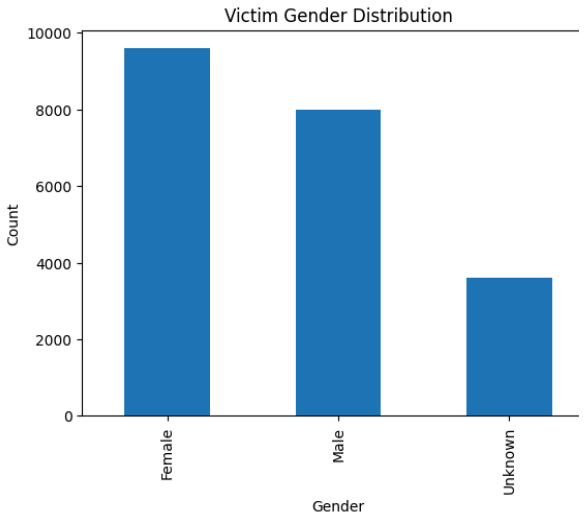
The raw dataset contained missing values and inconsistencies that required preprocessing before analysis:

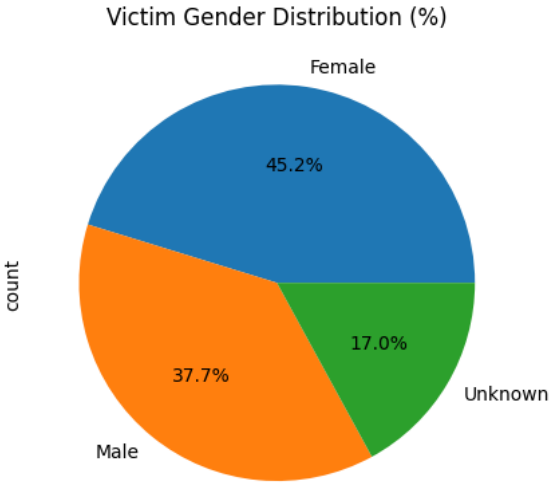
1. Handling Missing Data:
    - Categorical fields with missing values were filled with the category "Unknown".
    - Binary columns (e.g., isForcedLabour, IP_ControlAbuseKidnap) were standardized by replacing NaN with 0, ensuring a binary representation (1 for presence, 0 for absence).
2. Multi-Value Fields:
    - Fields like IP_Relation (e.g., "FamilyIntimatePartner;StrangerUnknownOther") and IP_ageBroad were expanded into binary flags for each category to facilitate analysis.
3. Data Type Adjustments:
    - Ensured categorical and numerical fields matched the expected formats, reducing memory usage and improving processing efficiency.

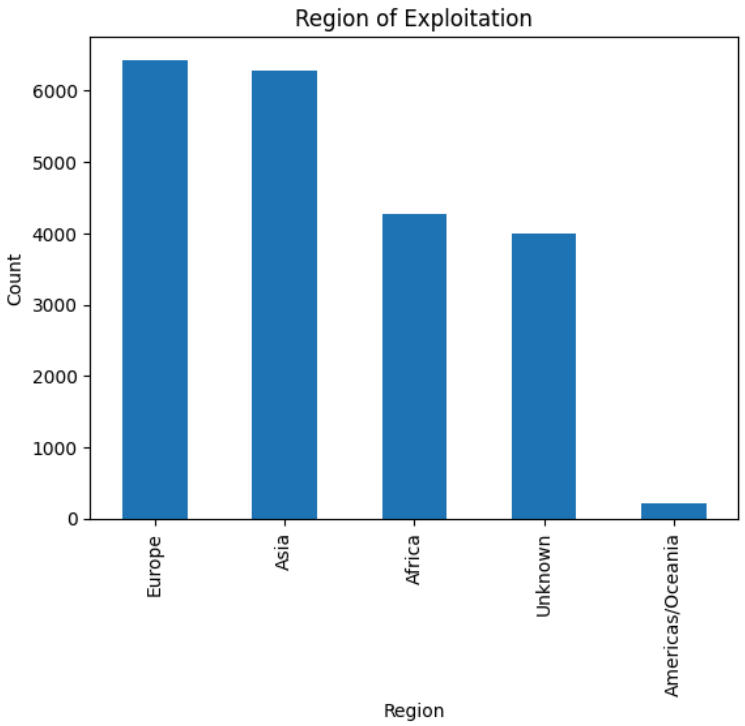
Exploratory Data Analysis

Initial analyses focused on understanding victim and perpetrator dynamics:

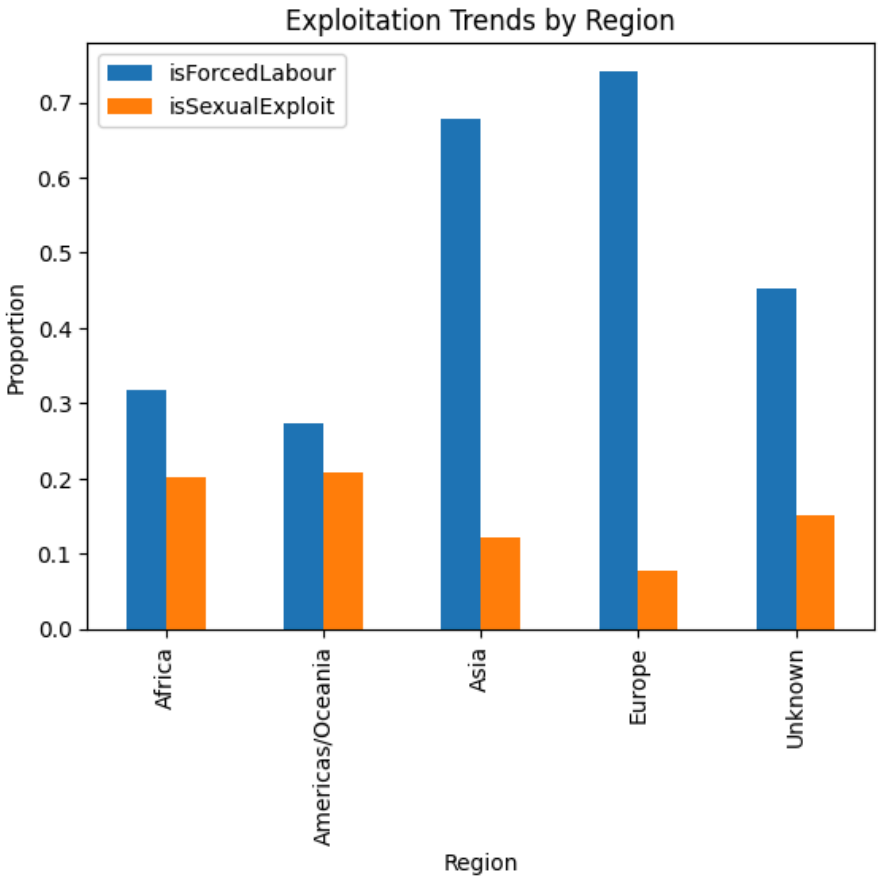
1. Demographics:
    - Gender distribution: Females accounted for the majority of victims in sexual exploitation, while males were more prevalent in forced labor.Age distribution: Minors were overrepresented in certain regions compared to adults.







2. Exploitation Types:
    - Forced labor was proportionally more dominant in Asia and Europe, and there was a higher proportion of sexual exploitation in Africa and the Americas. However, across the board, we can see that forced labor is more prevalent in all regions.

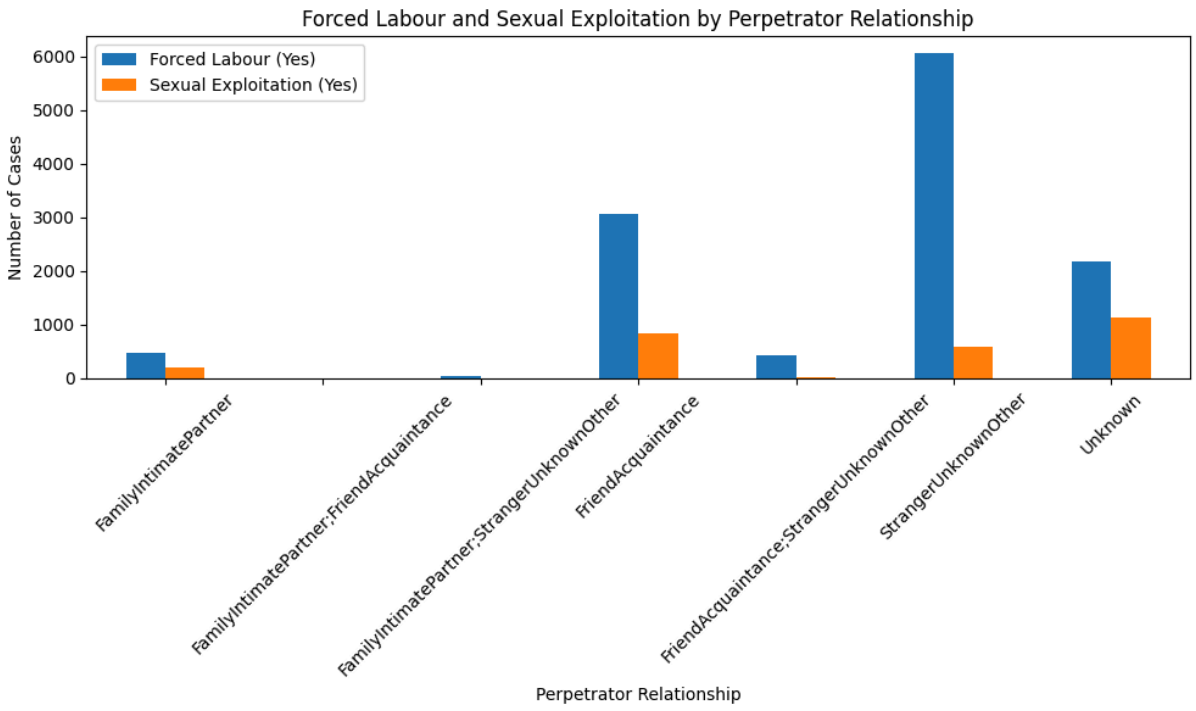


3. Victim-Perpetrator Relationships:
    - Family members and intimate partners were more commonly associated with sexual exploitation, whereas strangers played a significant role in forced labor.

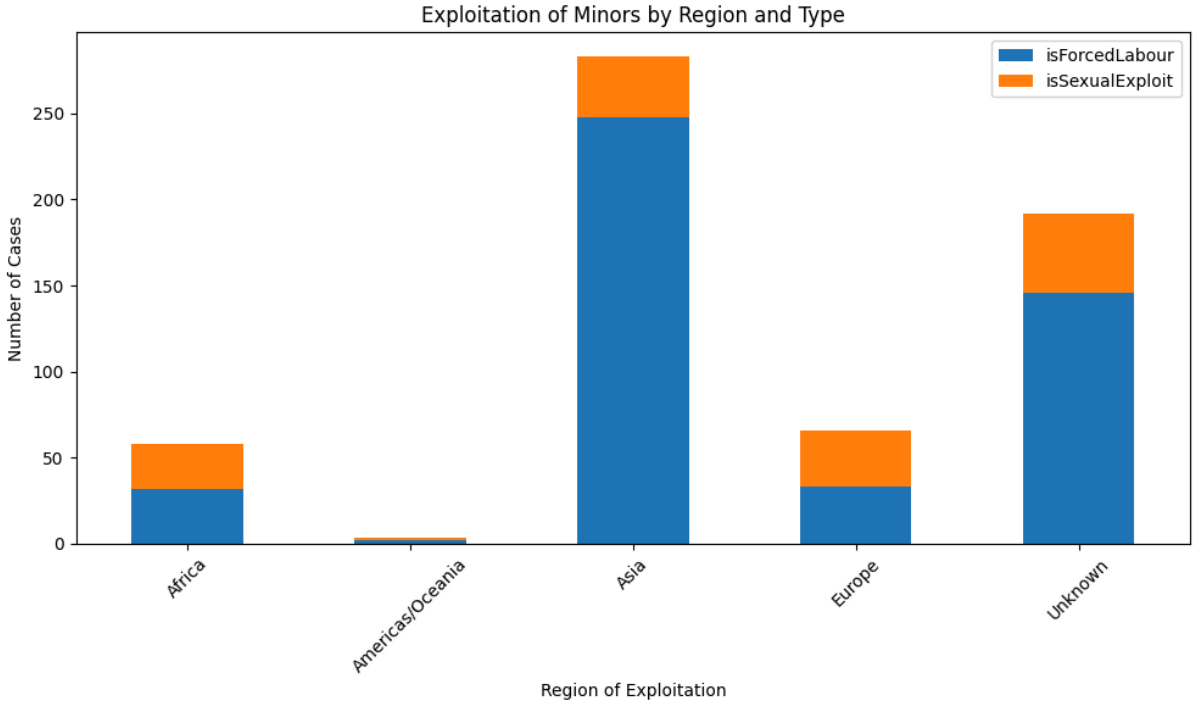
Forced Labor: Victim-Perpetrator Relationships		
Relation to Perpetrator	No	Yes
FamilyIntimatePartner	506	461
FamilyIntimatePartner;FriendAcquaintance	6	0
FamilyIntimatePartner;StrangerUnknownOther	35	38
FriendAcquaintance	1548	3069
FriendAcquaintance;StrangerUnknownOther	153	435

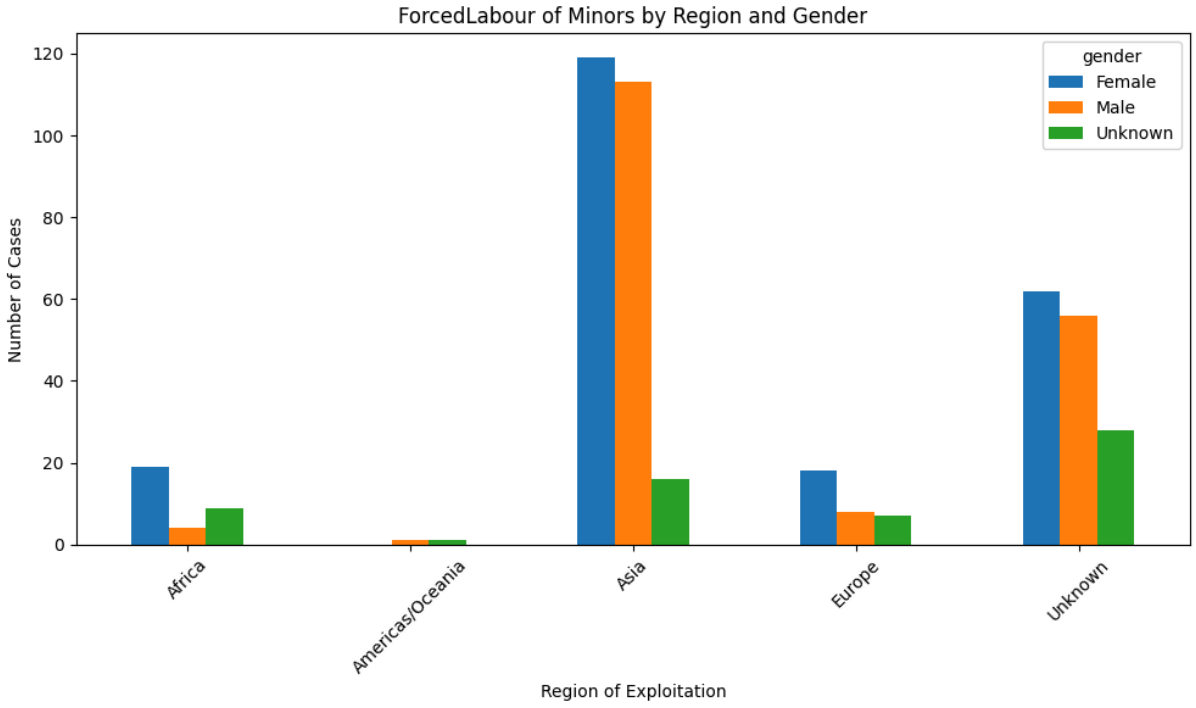
StrangerUnknownOther	2255	6066
Unknown	4444	2179

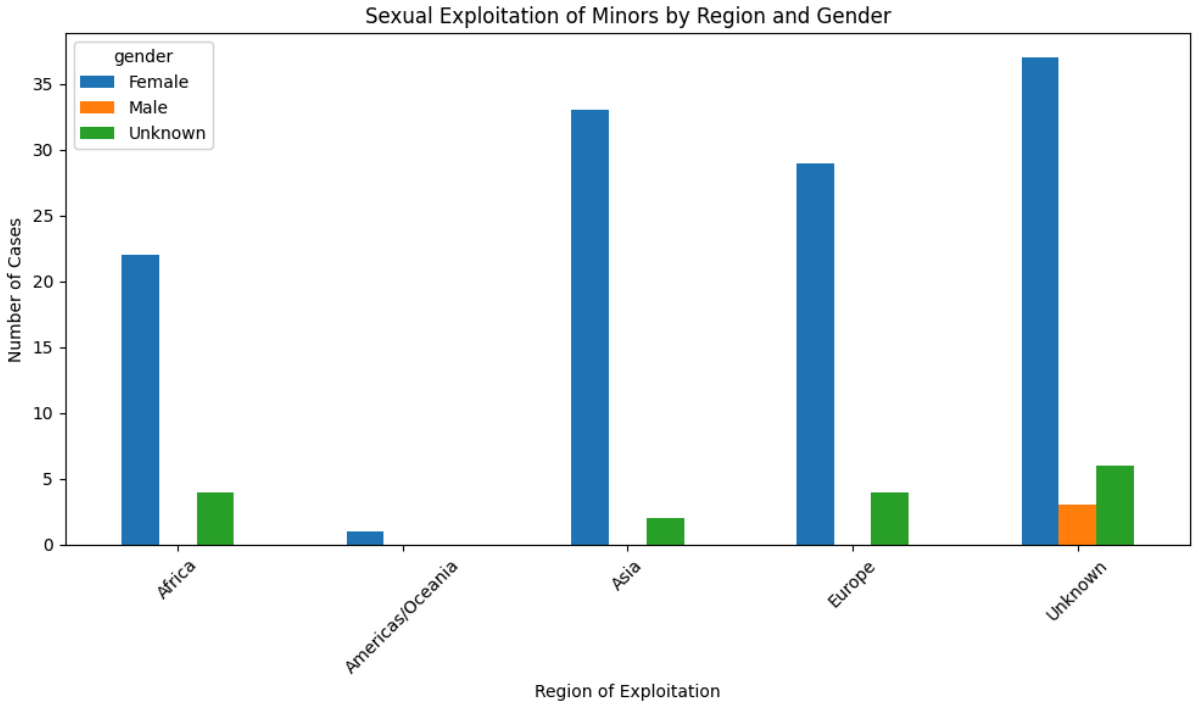
Sexual Exploitation: Victim-Perpetrator Relationships		
Relation to Perpetrator	No	Yes
FamilyIntimatePartner	767	200
FamilyIntimatePartner;FriendAcquaintance	6	0
FamilyIntimatePartner;StrangerUnknownOther	73	0
FriendAcquaintance	3790	827
FriendAcquaintance;StrangerUnknownOther	570	18
StrangerUnknownOther	7728	593
Unknown	5486	1137



4. Exploitation of Minors







We can observe that there is virtually few cases of sexual exploitation of male minors. With regard to forced labor, we can see that Asia has a relatively balanced distribution of male and female victims, while the remaining regions appear to have significantly more female victims.

Predictive Modeling

1. Classification Tasks:

    - Binary classification to predict isForcedLabour or isSexualExploit based on victim demographics, perpetrator roles, and regions.
    - Multiclass classification to predict UN_COO_Region (region of origin) using similar features.
2. Feature Engineering:
    - Interaction terms were created for combinations like gender × IP_Relation.
    - Numerical features were scaled using StandardScaler to improve model performance.

5. Results

Preliminary Insights

1. Binary Classification (Forced Labor vs. Sexual Exploitation):
    - Using Random Forest, we achieved an accuracy of 75% for predicting forced labor and 72% for sexual exploitation.
    - Key features:
        - Forced labor: Strongly associated with male victims and perpetrator roles like IP_ControlAbuseKidnap.
        - Sexual exploitation: Highly correlated with female victims and intimate partner relationships.
2. Region of Origin Prediction: Predicting what region a victim is from based on various features.
    - Logistic Regression achieved an overall accuracy of 65% for predicting UN_COO_Region (e.g., Africa, Asia, Europe).
    - Confusion Matrix Insights:
        - Strong accuracy for African and European origins.
        - Asia was the hardest to predict due to overlapping demographic and exploitation features.
3. Exploration of Patterns by Region:
    - Europe: Predominantly involved in labor exploitation cases.
    - Asia: High prevalence of forced labor.
    - Africa: Mixed exploitation types.

Challenges

- Imbalanced data for certain regions (e.g., low representation of Americas) affected model performance.
- Compound fields (e.g., IP_citizen_UNRegion) added complexity but also provided nuanced insights when expanded into binary flags.

6. Interpretation

7. Conclusion

8. Future Work

References

[1] About CTDC. Accessed: 2024-10-25. URL: https://www.ctdatacollaborative.org/page/about

[2] Global Synthetic Dataset. Accessed: 2024-10-25. URL: https://www.ctdatacollaborative.org/page/global-synthetic-dataset

[3] Victim-Perpetrator Synthetic Data Dashboard. Accessed: 2024-10-25. URL: https://www.ctdatacollaborative.org/dashboard/global-victim-perpetrator-synthetic-data-dashboard