AI-Enhanced Detection of Human Trafficking: Integrating Social Science Insights with Generative Artificial Intelligence for Homeland Security

Michael Backus
Department of Math and Computer Science
Fayetteville State University
mbackus@uncfsu.edu

Zach Delaney
Department of Math and Computer Science
Fayetteville State University
zdelaney@broncos.uncfsu.edu

Jonathankeith Murchison
Department of Math and Computer Science
Fayetteville State University
jmurchison1@broncos.uncfsu.edu

Shyamal Das
Homeland Security Program
Elizabeth City State University
sdas@ecsu.edu

Sambit Bhattacharya
Department of Math and Computer Science
Fayetteville State University
sbhattac@uncfsu.edu

***Abstract*—The detection of human trafficking remains a complex and urgent challenge, further compounded by the adaptive strategies employed by traffickers on online platforms. We propose a novel architectural model that integrates generative artificial intelligence (AI) with social science insights for the effective identification of human trafficking activities. The proposed software solution leverages advanced AI techniques, including computer vision, natural language processing (NLP), Large Language Models (LLMs) and deep neural networks (DNNs), to analyze web-based text-image data. Key components include biometric analysis via DeepFace and few-shot text classification, supported by synthetic data generation aligned with the Department of Homeland Security (DHS) Strategic Plan. To enhance performance, various vision-language models, detectors, and analytical methods were compared and integrated into a fusion-based tool capable of evaluating the likelihood that an online escort advertisement aligns with patterns indicative of human trafficking. Preliminary results demonstrate a notable accuracy rate exceeding 90% in identifying possible trafficking-related advertisements, underscoring the model's potential to significantly improve detection capabilities. Ongoing efforts focus on refining model precision and developing realistic synthetic image data to mitigate data scarcity challenges while maintaining ethical considerations. This interdisciplinary approach advances existing tools for homeland defense and law enforcement while prioritizing a victim-centered strategy. By leveraging AI-driven technologies, the proposed solution offers transformative potential for combating human trafficking and safeguarding vulnerable populations.**

***Index Terms*—Synthetic data, Facial recognition, Cross-modal retrieval, Biometric analysis, Human trafficking**

I. Introduction

The global challenge of human trafficking persists as a formidable threat to human security, impacting millions of lives around the world. The hidden and highly adaptive nature of sex trafficking networks (STNs) makes them particularly resistant to traditional forms of detection and intervention. In this complex landscape, innovative methodologies are essential not only for identifying trafficking activities but also for safeguarding potential and existing victims, many of whom are minors. Recognizing this urgent need, our research introduces an advanced architectural model that merges the capabilities of artificial intelligence (AI), social science, and homeland security strategies to enhance the detection and disruption of these clandestine operations.

The focus of our approach is the development of a software prototype that extends the Image Surveillance Assistant (ISA). This model harnesses the power of web-based technologies to amplify text-image capabilities, pivotal for uncovering and combating sex trafficking. By integrating modern AI techniques with insights from social science, our work aims to create a more nuanced and effective toolkit for homeland security and law enforcement. The interdisciplinary nature of this research facilitates a unique blend of computer vision, natural language processing (NLP), and deep neural networks (DNNs), targeting the complex interplay between text and visual data often used in trafficking advertisements. Prior efforts in text analysis have shown that unsupervised approaches can effectively uncover latent structure in escort ad corpora [15], which we extend here into a multimodal setting that leverages both NLP and computer vision [22].

A. Ethical Considerations and Bias in AI-Based Detection

The development of AI models for sex trafficking detection is fundamentally constrained by the sensitive nature of the underlying data. Unlike other machine learning domains where large-scale, annotated datasets are readily available, sex trafficking data is inherently scarce, fragmented, and ethically challenging to collect. The content often involves private, exploitative, or illegal material, requiring careful handling to avoid revictimization, data misuse, or privacy violations [17]. Legal restrictions, institutional barriers, and the need to protect survivors further limit researchers' access to real-world case data. Even when data is available—for example,

through law enforcement or NGOs—it is often anonymized, heavily redacted, or labeled inconsistently, complicating its use for supervised learning [18]. These limitations necessitate the exploration of alternatives such as synthetic data generation and unsupervised or semi-supervised learning approaches, while also demanding strict adherence to ethical guidelines and oversight. Fairness and contextual awareness remain critical to avoid disproportionately impacting vulnerable groups and reinforcing societal biases. As a result, the field must strike a careful balance between technological innovation and the moral responsibility to protect the individuals represented in the data.

B. Relevance and Operational Need

The effectiveness of dismantling criminal networks such as human trafficking hinges on a deep understanding of their operational dynamics. Law enforcement and security agencies must enhance their capabilities to gather and analyze relevant data from diverse sources, including open-source materials. Current AI tools, like those used in the DARPA Memex Program, assess risk by scoring online advertisements, such as those on escort websites [24]. However, these technologies lack the ability to perform cross-modal retrieval (See "Fig. 1"), integrating both text and image, to generate a likelihood or ambiguity score to determine if the advertisement involves trafficking or not.

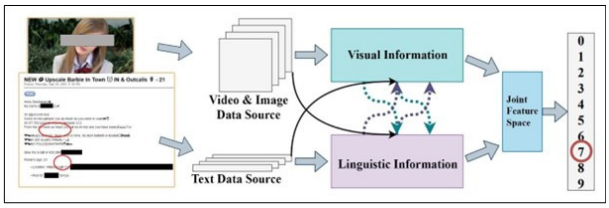

Fig. 1. Cross Modal Analysis on Sex Trafficking Advertisements

Addressing this deficiency, our research proposes a new architectural framework that enables the extension of existing web-based detection technologies. This framework will allow for the matching of text and image data to more accurately identify potential human trafficking instances on advertisement platforms. The integration of text and image analysis is crucial, particularly in domains where both explicit and implicit information, such as age, ethnicity, and location, are conveyed. Building on the work in the fashion sector (Gao et al, 2020) and prior research in semi-supervised learning for trafficking detection [10], we contend that in sex trafficking advertisements, both the overt and covert details are critical. For instance, accurately determining real age through image analysis could be pivotal in detecting underage victims of trafficking. Prior multimodal work, such as Tong et al. [11], demonstrates the value of deep models in jointly analyzing text and images for trafficking detection. By proposing an architecture like the Image Surveillance Assistant (ISA), this research aims to advance the field by developing tools that can effectively match and analyze text and image data from these advertisements, enhancing the detection and prevention of human trafficking.

II. Related Work & Existing Datasets

Sex trafficking (ST) has escalated notably over the past two decades, with a report from POLARIS indicating that 71% of human trafficking cases in the USA involve sex trafficking [1]. This issue has garnered significant attention since the early 21st century, and notably, reports to hotlines increased by over 45% during the 2020 pandemic [2]. While many studies such as those by Das et al. [3] and Shelley [7] provide theoretical and empirical insights into the drivers and mechanisms of ST, others like Kennedy [4] and Latonero [5] highlight the substantial profits driving innovations in this cross-national crime. Earlier machine learning efforts, such as those by Amin [13], focused on modeling and destabilizing trafficking networks, laying the foundation for more recent data-driven detection frameworks. Over the last decade, advancements in AI and machine learning have been pivotal in detecting ST cases. Latonero [6] discusses the challenges in confirming cases of Domestic Minor Sex Trafficking (DMST) through online ads, citing the use of keywords to increase the accuracy of identifying potential cases. However, these approaches have generally not included cross-modal retrieval that combines image analysis with textual data.

Despite the suspension of platforms like Craigslist and MySpace, many sites still facilitate sex trafficking. Some research has focused on community and population-level analyses to detect emerging trends but falls short in the identification of cases using both text and image data. Building on studies by Hundman et al. [9] and Kejriwal [8], our research proposes the development of a new AI framework, the Image Surveillance Assistant (ISA), aimed at enhancing the detection of sex trafficking through the integrated analysis of text and image data from online advertisements. This innovative approach seeks to address the limitations of current methodologies by providing a more comprehensive tool for detecting sex trafficking activities. Previous work on AI addresses the critical challenge of identifying perpetrators and victims who deliberately conceal their faces with masks or disguises in surveillance footage [25] .

This article outlines the development of a pioneering AI-driven model tailored to combat sex trafficking through enhanced text-image analysis. Our approach not only addresses the technical challenges inherent in detecting such complex criminal activities but also underscores the broader societal implications of employing AI in the fight against global security threats. Through continued evaluation and testing, we aim to deliver a deployment-ready prototype solution that can significantly bolster national and international efforts to dismantle sex trafficking networks and protect vulnerable populations.

The integration of artificial intelligence (AI) and social science insights in detecting human trafficking, particularly through online escort advertisements, has gained traction in recent years. This literature review focuses on the available

trafficking datasets, notably the Trafficking-10k dataset, the use of generative AI to create synthetic datasets, and the contributions of social science in identifying patterns that may indicate trafficking. Several efforts have explored leveraging publicly available data to identify trafficking patterns using machine learning techniques [12].

The Trafficking-10k dataset is a significant resource for researchers aiming to develop machine learning models for detecting human trafficking in online advertisements. This dataset comprises 10,000 escort ads labeled by anti-trafficking experts on a scale from "very unlikely" to "very likely" to contain trafficking victims. Researchers have leveraged this dataset to train various AI models, including ordinal regression neural networks, which have shown improved performance in identifying potentially trafficked individuals in online ads. However, the dataset is not without limitations. Access constraints and privacy issues hinder the availability of comprehensive data, and the inherent biases in the dataset may lead to misinterpretations of the ads. Additionally, the reliance on expert labeling can introduce subjectivity, which may affect the consistency and reliability of the dataset. Our team was unable to access the Trafficking-10k dataset directly due to prior approval requirements.

Generative AI has emerged as a promising avenue for creating synthetic datasets that can augment existing trafficking datasets. By generating realistic text ads and images, researchers can address the limitations of current datasets, such as privacy concerns and the scarcity of labeled data [16]. This approach allows for the creation of diverse scenarios that reflect the complexities of human trafficking, thereby enhancing the robustness of machine learning models. Prior work has demonstrated the potential of generative AI in producing synthetic data that mimics the linguistic patterns found in real escort advertisements, which can be crucial for training models to recognize nuanced indicators of trafficking.

Social science insights play a critical role in understanding the language and symbols used in online escort advertisements. Studies have identified specific lexicons, emojis, and other indicators that may suggest the presence of trafficking. For instance, the use of emoticons in ads has been explored as a potential indicator of trafficking, highlighting the evolving nature of language in this context. Furthermore, research has shown that certain phrases and terms are more prevalent in ads associated with trafficking, which can inform the development of more effective detection algorithms. By integrating these social science findings with AI methodologies, researchers can enhance the accuracy of trafficking detection systems.

Despite the advancements in using trafficking datasets and generative AI, challenges remain. The ethical implications of using synthetic data, particularly in sensitive contexts like human trafficking, must be carefully considered. The potential for misrepresentation or the creation of harmful stereotypes through synthetic data generation poses significant risks. Moreover, the dynamic nature of online advertising necessitates continuous updates to datasets and models to ensure their relevance and effectiveness in combating trafficking.

III. Methodology

Our methodology integrates synthetic data generation, multimodal analysis, and AI-driven classification to detect human trafficking indicators in online advertisements. The system is composed of three primary components: synthetic dataset creation, model-based detection, and a fusion framework for final scoring.

A. Generating Synthetic Data for Trafficking Ad Text

To simulate realistic examples of trafficking-related advertisements, we constructed a synthetic dataset designed to resemble real-world online ads potentially associated with human trafficking activity.

TABLE I
Synthetic Text Advertisement Dataset Composition

Category	Number of Ads
Trafficking-related Ads	500
Non-trafficking Ads	500
Total	**1000**

This process began by collecting publicly available data and leveraging open-source datasets containing linguistically stylized trafficking-related advertisements (See TABLE I). These examples often featured indirect language, coded slang, or euphemistic phrasing—elements crucial for training a generative model capable of producing realistic synthetic content.

The generation of synthetic data for trafficking advertisements enables us to automate the creation of various ads, which can then be utilized in our detection methods and models [16].

Initially, we explored synthetic text generation using OpenAI's ChatGPT. However, this approach was hindered by content filtering mechanisms that are commonly implemented across several AI tools for ethical reasons, particularly concerning sensitive topics like trafficking.

For synthetic text ad generation, we employed a multi-tiered process assisted by a large language model (LLM), specifically mlabonne/NeuralDaredevil-8B-abliterated. This model was chosen for its lack of safety checks, allowing it to generate sensitive data that other LLMs would typically block. The process began by selecting random attributes such as hair color, eye color, build, and height. These limited variables were deliberately chosen to standardize and categorize the generated data, which also helped minimize hallucinations during text creation.

Structured templates were then populated with these random attributes and passed to the LLM to generate synthetic ads, guided by a set of sample ads. This approach ensured consistency in the generated content while providing a reliable framework for categorizing and analyzing potential trafficking indicators. The same method was adapted to create non-trafficking ads, replacing attributes with job descriptions and keywords to generate realistic but harmless ad text (See "Fig. 2").

Additionally, we employed a similar method for emoji generation, leveraging the LLM's contextual understanding to

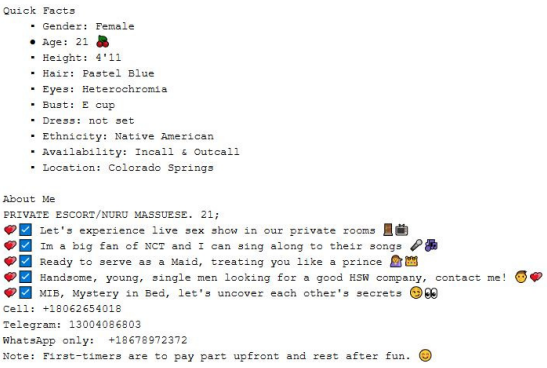

Fig. 2. Example synthetic ad

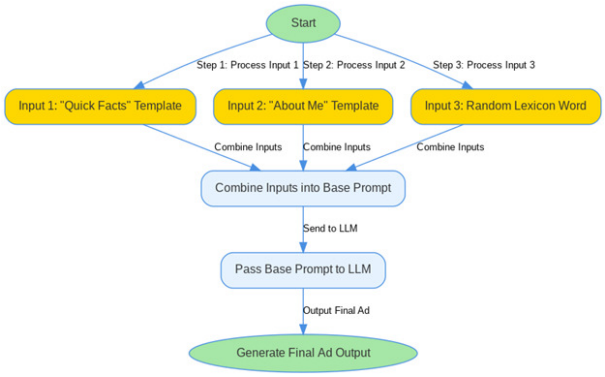

Fig. 3. Synthetic Data Generation Workflow

associate appropriate emojis with text content. This enabled us to analyze how emojis are used within certain types of ads, providing further insight into subtle indicators that might be present in trafficking-related material. This approach is consistent with prior research that combines NLP and visual analysis to detect illicit messaging patterns across social platforms [22].

To enhance the process of synthetic text ad generation, we utilized two dedicated programs: traffick_ad_text_gen.py and non_traffick_ad_text_gen.py. These programs served as the backbone for creating a standardized dataset of trafficking-related and non-trafficking ads. (See "Fig. 3").

traffick_ad_text_gen.py: This program specifically focuses on generating trafficking-related text ads. It integrates the multi-tiered process described, using the mlabonne/NeuralDaredevil-8B-abliterated model to populate structured templates with randomly selected attributes, such as hair color, eye color, build, and height. By leveraging the model's capacity for unrestricted content generation, this tool produces synthetic ads aligned with the characteristics and subtle indicators commonly associated with trafficking scenarios.

non_traffick_ad_text_gen.py: This counterpart program adapts the same methodology but replaces the trafficking-related attributes with job descriptions, keywords, and neutral attributes.

The goal is to generate realistic but entirely benign text ads, offering a reliable comparative framework for distinguishing between trafficking and non-trafficking ad patterns.

The structured templates, consistent attribute selection, and model-guided generation process implemented in these programs provide a robust mechanism for synthesizing diverse datasets. Additionally, the incorporation of emoji generation, facilitated by the LLM's contextual capabilities, adds depth to the dataset, enabling nuanced analyses of how emojis are utilized in various ad types.

These programs not only standardize and streamline the generation of text ad datasets but also facilitate their integration into analytical workflows, offering valuable tools for identifying and studying potential trafficking indicators.

B. Synthetic Image Generation

To complement synthetic text advertisements, we employed text-to-image generation techniques to simulate the visual environments described in trafficking-related ads. Using models such as Stable Diffusion, DALL-E, and MidJourney, we attempted to generate images based on descriptive text attributes—e.g., room type, lighting, and setting. However, these platforms imposed content restrictions, particularly on imagery involving underage appearances, bruises, or suggestive clothing, limiting their applicability.

To address these limitations, we explored alternative tools with fewer restrictions, such as SexyAI, and evaluated locally hosted solutions like Foocus (a Gradio-based image generation platform inspired by Stable Diffusion and MidJourney). While promising, hosting complexities led us to continue exploring other flexible solutions.

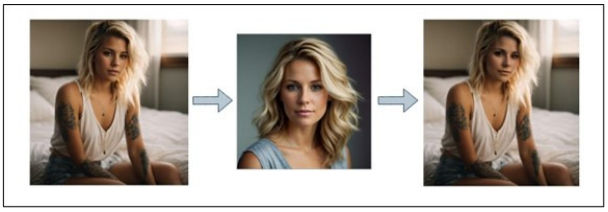

Fig. 4. Example of synthetic image generated to simulate trafficking-related visual cues

To achieve greater control over visual demographics, we created a diverse portrait library and used DeepFace for facial attribute analysis. DeepFace, particularly with the RetinaFace backend, provided reliable age, gender, and race estimations. Since many image generators poorly replicate age specifications, we used DeepFace to screen and select portraits that closely matched the target demographic. We then applied face-swapping techniques to align these portraits with the advertisement context, enabling precise control over perceived age (See "Fig. 4").

This pipeline allowed us to create synthetic images that are both ethically compliant and suitable for training detection models. While tools like LeonardoAI were also considered,

ethical constraints and restrictive content filters prevented their practical use for trafficking-specific scenarios.

By combining image generation with DeepFace-based validation and augmentation, we established a scalable, controlled method for producing synthetic training images. These visuals support downstream detection tasks by enriching multimodal datasets with consistent and demographically realistic image content.

C. Detection of Trafficking based on Text and Images

Text processing involves automated analysis of electronic text, crucial in modern data analysis. We plan to use an advanced AI text generator, leveraging natural language processing (NLP) and machine learning, to produce human-like text from prompts. By training this AI on a diverse corpus of text data, including books, articles, and websites, we aim to enhance its understanding of language patterns and syntax for generating accurate phrases. Our goal is to use these phrases to identify high-risk advertisements. Prior research has shown that unsupervised text template matching can be effective in identifying structured patterns in escort ads without labeled training data [14], supporting the value of phrase-level analysis in trafficking detection. Recent work also highlights how synthetic text generation using machine learning can expand training data when labeled examples are limited [16]. Although we initially considered ChatGPT, implementation issues led us to explore ”jailbreaking” to achieve the desired outputs.

In addition to generative models, we fine-tuned a multilingual BERT model to support classification and detection tasks across multiple languages. BERT, originally introduced by Devlin et al. [19], has become a foundational model in natural language processing, enabling a wide range of tasks such as sentiment analysis, named entity recognition, and document classification [20]. Its application to human trafficking detection has also been demonstrated in recent work focused on analyzing escort advertisements using language models [21]. This model allows us to analyze ad content not only in English but also in various other languages commonly found in trafficking-related posts. Its ability to generalize linguistic patterns across languages enhances our system's ability to flag high-risk content regardless of the ad's original language, aligning with broader efforts to apply machine learning across diverse digital platforms for human trafficking detection [15].

Image processing for age, gender, and race detection is complex due to human appearance variability and nuanced differences. Age detection is challenging due to subtle, gradual facial changes, while gender detection faces difficulties from diverse gender expressions. Race detection is complicated by diverse skin tones, facial structures, and potential biases from training data. These tasks require sophisticated models trained on extensive datasets to ensure accuracy, yet misclassifications and biases persist, highlighting the need for continued advancements.

To address these challenges, we used the DeepFace library for detecting age, race, and gender data. Among various backends, RetinaFace, based on MXNet and developed by InsightFace, provided the most robust results.

D. Auxiliary Detection Models: NSFW_MODEL

At the early stages of this study, we utilized the NSFW Model available at to analyze and classify images for potential NSFW (Not Safe For Work) content. The objective was to determine whether this approach could aid in identifying imagery associated with trafficking by filtering out benign or harmless images. The model demonstrated effectiveness in categorizing images and provided a useful baseline for identifying content with higher sensitivity to explicit imagery. However, during the analysis, we observed a notable bias towards categorizing images of females as NSFW, which limited its applicability for this specific use case. This bias highlights the challenges of generalizing pretrained models for sensitive applications and underscored the need for more targeted tools to address the unique requirements of trafficking detection.

E. Fusion and Detection Capabilities

To estimate the likelihood of human trafficking in online advertisements, we developed a multimodal fusion framework that integrates text and image analysis. This approach combines outputs from several tools: DeepFace for visual attributes (age, race, emotion), a multilingual BERT model for text classification.

The process begins with a synthetic dataset consisting of text-based ads, each paired with a corresponding synthetic image. For text, we fine-tuned a multilingual BERT model capable of detecting high-risk content across multiple languages. For images, DeepFace and Amazon Rekognition were used to extract age estimates, emotional state (e.g., fear, sadness), and contextual scene attributes.

These distinct signals—linguistic, visual, emotional, and environmental—were aggregated using a decision tree–based fusion architecture. Each input contributed a weighted score, culminating in a final trafficking likelihood score normalized between 0 and 1. This multimodal approach improves detection accuracy while enhancing contextual understanding, aligning with recent work on vision-language modeling for sensitive content detection [23].

IV. Results

Our results are broken down between many smaller tasks and requirements. Using synthetic data, we were able to leverage open-source libraries and hope to build a fusion application using various parts of our results (See “Fig. 5”).

A. Synthetic Image and Text generation Results

Given the real data our team is modeling is either extremely sensitive or classified as the real data may either contain issues with privacy with of-age victims, or potentially child pornography (CP) of trafficked, underage victims, this domain of research is fraught with potential ethics or even legal risks. Beyond the possibility of our methods being able to accurately model the real data in a safe manner, there is a

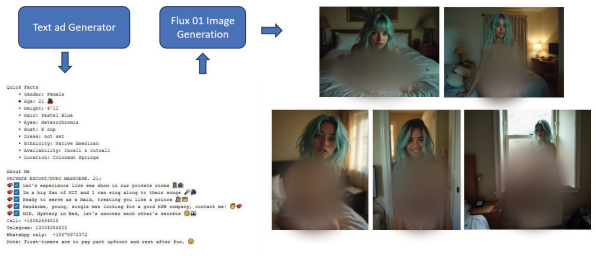

Fig. 5. Workflow from synthetic ad to synthetic image generation

limit to the synthetic dataset the team can and should be allowed to produce. In one respect the generative AI tools themselves perform this boundary check externally, and certain words are filtered out of such requests. For example, the generative AI image tools like MidJourney and DALL-E forbid words like "young" and "prostitute" in the context of such an image. Creating images where exploitation of humans is either forbidden/against the Terms of Service (ToS) of most of these image generation tools.

We performed analysis on 170 initial useable synthetic images and text using AI sources that allowed us to test various libraries on synthetic data, confirming that synthetic data will be a valid tool in building a fusion program to help detect trafficking.

B. Synthetic Text Analysis Results

Using a multilingual BERT model, we fine-tuned the classifier on our synthetic dataset to identify trafficking-related text patterns (See TABLE II). The model demonstrated strong performance in detecting high-risk indicators, even when the ads were written in multiple languages. Fine-tuning on synthetic text allowed us to simulate varied linguistic styles, tones, and contexts often seen in real-world trafficking scenarios.

The model achieved an F1 score of 0.91 on our validation set, showing its effectiveness in distinguishing between trafficking-related and benign ads. Key features contributing to high-risk classification included the presence of coded language, frequent geographic references, and specific terms correlated with underage or exploitative contexts.

The multilingual capacity of the model enabled it to generalize across English, Spanish, and other commonly used languages in trafficking advertisements, significantly improving cross-lingual detection accuracy compared to a monolingual baseline. These findings support the value of synthetic training data combined with multilingual modeling in building scalable and language-agnostic detection systems.

C. Synthetic Image Analysis Results using Deepface

Using the open-source library DeepFace we were able to accurately test our synthetically generated images. Since we passed a text-based prompt to the image generator we can test DeepFaces capabilities in determining things like race and gender. Based on an initial analysis DeepFace was able to successfully predict gender of synthetic images 93% of the time. While predicting proper race of synthetic images 79% of the time. These help us to adapt to the library's biases. TABLE III details other experimental results on this project.

TABLE II
BERT MODEL CONFUSION MATRIX ON SYNTHETIC TEST SET

Predicted	Trafficking	Non-Trafficking
Trafficking	93	7
Non-Trafficking	3	97

Note: The model was trained on 400 synthetic trafficking and 400 non-trafficking examples, and evaluated on a test set of 100 trafficking and 100 non-trafficking examples. Overall accuracy was 95%. Slight class imbalance is visible in the error distribution.

TABLE III
SUMMARY OF KEY EXPERIMENTAL RESULTS

Task	Result
Synthetic Data Generation	170 usable synthetic ads/images created to test pipeline.
Text Classification (BERT)	Trained on 400 trafficking / 400 non-trafficking examples. Test set: 100 trafficking / 100 non-trafficking. Accuracy 95%
Image Analysis (DeepFace)	Gender detection accuracy: 93%; race detection accuracy: 79%. Age detection reliable for 18–55 age range.
Fusion Program Testing	Applied to small, unseen synthetic subset; correctly predicted all examples, confirming feasibility of scoring approach.

A noted issue with synthetic image generation and age is most AI's do not have a solid grasp of age-based numbers. Much like humans determining age, it can be very subjective between many age groups and because of this it is hard to determine whether the age of a synthetically generated image is accurate. To fully assess DeepFace' age detection we used the UTKFace dataset to determine age accuracy. This allowed us to test multiple backends to pick the best one. What we found was the DeepFace library is very capable of determining age between the ranges of 18-55 but struggled with ages outside of the 18-55 range. We determined that a broad range of ages is much more practical for our AI based detection.

D. Fusion Program Analysis of Synthetic Data

After generating synthetic text and corresponding synthetic images, we took our synthetic dataset and fine-tuned a multilingual BERT model to evaluate indicators of trafficking. One component of the evaluation involved a scoring matrix based on detected attributes from the synthetic images and text. Subjects detected as being under the age of 18 contributed significantly to the trafficking likelihood score. Emotional analysis was also conducted—negative emotions such as fear, sadness, disgust, and anger increased the overall score. Additionally, the original synthetic ad text was analyzed for the presence of trafficking-related keywords. These combined factors produced a normalized trafficking score between 0 and 1, as illustrated in Fig. 6. We note that we only tested the fusion program on a small subset of previously unseen synthetic ads

and images; although limited in scope, it correctly predicted all examples in this subset.

Our research highlights the potential and challenges of using a multimodal approach to detect human trafficking advertisements. By combining lexicon-based detection with image analysis through DeepFace and AWS Rekognition, and augmenting this with a fine-tuned multilingual BERT model, we demonstrated that a scoring system integrating textual and visual cues can effectively identify trafficking indicators (See "Fig. 6"). The BERT model enhanced the text classification pipeline by enabling detection of high-risk content in both English and non-English languages, significantly improving the generalization of our detection system. The image analysis, which extracted emotion, age, and race data, also showed promising results, with less than a 5% variance between detections and false positives. However, our study also revealed critical limitations of using synthetic data in this context.

V. Conclusion

The research presented outlines an innovative interdisciplinary approach to combating human trafficking by integrating advanced artificial intelligence techniques with insights from social science and homeland security. Through the development of a multimodal framework that fuses text and image analysis, this study demonstrates how synthetic data generation, natural language processing, and biometric analysis can collectively enhance the detection of trafficking-related indicators in online environments. By achieving high detection accuracy and validating the feasibility of fusion-based methods, this work contributes a scalable foundation for future AI-driven systems that support law enforcement and victim-centered investigations. One significant limitation is the emotional authenticity gap. Synthetic datasets often fail to capture the nuanced emotional expressions seen in real trafficking victims, leading to potential blind spots in image-based emotion detection. While fear and sadness were detected at similar rates across both our trafficked and non-trafficked datasets, this uniformity suggests that synthetic data may not accurately reflect the emotional variations present in real-world trafficking ads.

Our initial ad generation system relied on a lexicon of only 84 words or phrases, from which three words were randomly selected to populate a template prompt. Although this method produced workable ads, the limited vocabulary restricted the model's creativity, and the selected words were not always contextually related. This highlights a future research direction: expanding the lexicon and ensuring semantic cohesion between selected terms. Our fusion program's performance was heavily dependent on the size and diversity of this lexicon. Future research should therefore focus on incorporating context-aware word selection mechanisms—possibly guided by transformer models like BERT—to significantly enhance the quality and applicability of the generated ads.

To further improve multilingual detection, our approach included a fine-tuned multilingual BERT model capable of identifying trafficking indicators in ads written in various languages. This capability proved critical for extending the reach of our detection system beyond English-only datasets, aligning with the global nature of human trafficking.

$$S_{\text{trafficking}} = \min(S_{\text{keywords}} + S_{\text{BERT}} + S_{\text{age}} + S_{\text{emotion}}, 1.0)$$

Keywords:

$$S_{\text{keywords}} = \min(n_{\text{keywords}} \times 0.05, 0.6)$$

BERT:

$$S_{\text{BERT}} = P_{\text{BERT, trafficking}} \times 0.5$$

Age:

$$S_{\text{age}} = \begin{cases} 1.0, & \text{if age} < 18 \\ 0.2, & \text{if age} \leq 20 \\ 0, & \text{otherwise} \end{cases}$$

Negative emotion > 5%:

$$S_{\text{emotion}} = \min(n_{\text{negative emotions}} \times 0.1, 0.2)$$

Combined with Image:

$$S_{\text{profile}} = \min(S_{\text{text}} + S_{\text{images}}, 1.0)$$

$$S_{\text{text}} = S_{\text{keywords}} + S_{\text{BERT}}$$

N = number of images

$$S_{\text{images}} = \begin{cases} \frac{\sum S_{\text{image}}}{N} \times 2, & \text{if } N > 1 \\ \frac{\sum S_{\text{image}}}{N}, & \text{otherwise} \end{cases}$$

Fig. 6. Trafficking score Generation

While we were successful in generating ads with emojis and most of those emojis were contextually accurate to the ad, we had a limited emoji set and relied on the LLM's contextual understanding of the emojis. While we provided some emojis in the template, a better emoji dictionary that aligns emojis with trafficking-related meanings would be beneficial in future research, especially in multilingual contexts where cultural interpretations of emojis may vary.

Despite efforts to limit hallucinations in our text ad generation, we still experienced significant hallucinations from the model. The templated generation approach helped constrain outputs and made parsing easier, yet we still observed a 15% failure rate. Failure rate was defined as generated ads that deviated from the expected structure, requiring additional reruns of the program. This is significant, as the average time to produce unique, clean ads was approximately 10 minutes per batch. For operational deployment of the techniques demonstrated in our research, human oversight remains essential. Derived model predictions should serve as decision-support tools rather than automated indicators of trafficking.

Another limitation lies in age representation. Although AWS Rekognition performed well on age verification datasets in prior experiments, our synthetic dataset did not adequately represent younger individuals, restricting the algorithm's ability to leverage age as a meaningful trafficking indicator. This is a notable constraint, as statistical data indicate that younger victims are more likely to be trafficked. Without sufficient underage representation, our detection model could not fully realize its potential in identifying high-risk cases.

Additionally, demographic constraints posed a challenge. Race, while potentially relevant to trafficking detection, was not effectively incorporated into the model due to the limitations of synthetic data diversity. However, race may be a

stronger indicator in different regions and could be weighted for the needs of a particular area. General detection of race would be more useful if applied in areas with a higher rate of trafficking among certain racial groups. Incorporating such regional sensitivities could strengthen the model's accuracy in real-world applications.

Despite these challenges, our findings suggest that a well-integrated multimodal approach—combining lexicon analysis, a multilingual BERT model for cross-lingual text detection, and facial attribute detection—significantly enhances trafficking ad detection compared to relying on any single modality. Moving forward, refining these methods with real-world data could mitigate the limitations observed with synthetic datasets, improving detection accuracy and validating the robustness of our scoring system.

In conclusion, while synthetic data offers a valuable testing ground, it cannot fully replicate the complexities of real trafficking advertisements. Yet, its potential remains hopeful—as a scalable, controlled environment for initial testing and algorithm refinement. Future work should prioritize incorporating authentic datasets to train and validate detection models, ensuring they perform reliably in real-world applications. By addressing the identified gaps, we believe this multimodal approach holds significant promise for aiding law enforcement and organizations in the fight against human trafficking.

The research presented outlines an innovative approach to combating sex trafficking by integrating advanced artificial intelligence techniques with social science insights and homeland security strategies. At the forefront of our technological arsenal is DeepFace, a cutting-edge facial attribute analyzer that enables the precise prediction of age, race, and gender characteristics. This technology not only enhances our current capabilities but also holds promise for future efforts to narrow down potential trafficking victims through rendered age progressions. By developing sophisticated software, we aim to significantly enhance the detection and disruption of trafficking activities, all while maintaining a steadfast victim-centered approach.

Looking ahead, we are exploring future advancements that include the automation of text and image classification and the synthetic generation of data to improve the training of future models. We are also investigating innovative methods to bypass content filtering, such as the introduction of xAI Grok-1 image and text generation, which we believe will allow us to surmount some of our most formidable challenges. Through these efforts, we aim to remain at the cutting edge of technological solutions in the ongoing battle against sex trafficking.

Acknowledgement

This work was supported in part by the Criminal Investigations and Network Analysis (CINA) Center at George Mason University, with prime funding from the U.S. Department of Homeland Security. The funded project is titled "An Architectural Model for Web-Based Technologies to Enhance Text-Image Capabilities in Detecting Sex Trafficking Cases."

References

[1] POLARIS, "Sex trafficking in or from Latin America," 2018. [Online]. Available:

[2] POLARIS, "Sexual exploitation during the pandemic," 2021. [Online]. Available:

[3] S. Das, L. Eargle, and A. Esmail, "Cross-national ST network in developing countries: A theoretical overture using global commodity chain approach," *J. Race, Sex and Class*, vol. 18, no. 1–2, pp. 230–253, 2011.

[4] E. Kennedy, "Predictive patterns of ST online," Senior Honors Thesis, Carnegie Mellon University, 2012. [Online]. Available:

[5] M. Latonero, "Human trafficking online: The role of social networking sites and online classifieds," *SSRN Electronic Journal*, 2011. doi:10.2139/ssrn.2045851

[6] M. Latonero, "The rise of mobile and the diffusion of technology-facilitated trafficking," Center for Communication Leadership & Policy, 2012. [Online]. Available:

[7] L. Shelley, "Human trafficking and trafficking into Europe: A comparative perspective," Migration Policy Institute, 2014. [Online].

[8] M. Kejriwal, "A meta-engine for building domain-specific search engines," *Software Impacts*, vol. 7, 100052, 2021.

[9] L. Whitney, J. Thomas, and M. Kejriwal, "Generative Techniques for Human Trafficking Detection in Escort Advertisements," *Proceedings of the AAAI Workshop on AI for Social Good*, 2018.

[10] H. Alvari, P. Shakarian, and J. K. Snyder, "Semi-supervised learning for detecting human trafficking," *Security Informatics*, vol. 6, no. 1, pp. 1–14, 2017.

[11] E. Tong *et al.*, "Combating Human Trafficking with Multimodal Deep Models," *arXiv preprint arXiv:1705.02735*, 2017.

[12] A. Dubrawski, K. Miller, M. Barnes, B. Boecking, and E. Kennedy, "Leveraging publicly available data to discern patterns of human-trafficking activity," *Journal of Human Trafficking*, vol. 1, no. 1, pp. 65–85, 2015.

[13] S. Amin, "A Step Towards Modeling and Destabilizing Human Trafficking Networks Using Machine Learning Methods," in *Proc. AAAI Spring Symp. Artificial Intelligence for Development*, 2010.

[14] L. Li *et al.*, "Detection and characterization of human trafficking networks using unsupervised scalable text template matching," in *Proc. 2018 IEEE Int. Conf. Big Data*, pp. 3111–3120, 2018.

[15] National Academies of Sciences, Engineering, and Medicine, *Facial Recognition Technology: Current Capabilities, Future Prospects, and Governance*, Washington, DC: The National Academies Press, 2024.

[16] Y. Lu *et al.*, "Machine Learning for Synthetic Data Generation: A Review," *arXiv preprint arXiv:2302.04062*, 2023.

[17] R. Denton *et al.*, "Ethical Tensions in Applications of AI for Addressing Human Trafficking: A Human Rights Perspective," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, 2022.

[18] K. Hundman, T. Gowda, M. Kejriwal, and B. Boecking, "Always Lurking: Understanding and Mitigating Bias in Online Human Trafficking Detection," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 137–143, 2018.

[19] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[20] M. V. Koroteev, "BERT: A Review of Applications in Natural Language Processing and Understanding," *arXiv preprint arXiv:2103.11943*, 2021.

[21] J. Zhu, L. Li, and C. Jones, "Identification and Detection of Human Trafficking Using Language Models," in *2019 European Intelligence and Security Informatics Conference (EISIC)*, pp. 24–31, 2019.

[22] S. Granizo, A. Caraguay, L. Lopez, and M. Hernandez-Alvarez, "Detection of possible illicit messages using natural language processing and computer vision on Twitter and linked websites," in *IEEE Access*, vol. 8, pp. 44534–44546, 2020.

[23] X. Han, X. Chen, Y. Song, Y. Hu, and Y. Zhao, "Multimodal Fusion and Vision-Language Models: A Survey for Robot Vision," *arXiv preprint arXiv:2504.02477*, 2025.

[24] MIT Lincoln Laboratory, "Turning technology against human traffickers," MIT News, May 6, 2021. [Online].

[25] Hong, Qi, et al. "Masked face recognition with identification association." *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2020.