Wildlife Product Trading in Online Social Networks:
A Case Study on Ivory-Related Product Sales Promotion Posts

Guanyi Mou[*], Yun Yue[*], Kyumin Lee, Ziming Zhang

Worcester Polytechnic Institute
100 Institute Road, Worcester, MA, 01609
{gmou, yyue, kmlee, zzhang15}@wpi.edu

Abstract

Wildlife trafficking (WLT) has evolved into a pressing global concern, as traffickers increasingly utilize online platforms such as e-commerce websites and social networks to expand their illicit trade. This paper addresses the pivotal challenge of detecting and recognizing promotional behaviors related to the sale of wildlife products within online social networks—a critical step in combating these environmentally detrimental activities. To confront these illicit operations effectively, our research undertakes the following key initiatives: 1. Data Collection and Labeling: We employ a network-based approach to gather a scalable dataset pertaining to wildlife product trading. Through a human-in-the-loop machine learning process, this dataset is meticulously labeled, distinguishing between positive class samples containing wildlife product selling posts and hard-negatives representing regular posts misclassified as potential WLT posts, subsequently rectified by human annotators. 2. Machine Learning Framework Development: We present a robust framework that benchmarks machine learning results on the collected dataset. This framework autonomously identifies suspicious wildlife selling posts and accounts, effectively harnessing the multi-modal nature of online social networks. 3. In-depth Analysis of Trading Behaviors: Our research delves into a comprehensive analysis of trading posts, illuminating the systematic and organized selling behaviors prevalent in the current landscape. By providing detailed insights into the nature of these behaviors, we contribute valuable information for understanding and countering illegal wildlife product trading. Moreover, we emphasize our commitment to openness and collaboration by making our code and dataset openly available, thereby fostering cooperative efforts towards the development of more effective strategies in combating illegal wildlife trafficking.

1 Introduction

Wildlife trafficking, defined as "the poaching or other taking of protected or managed species and the illegal trade in wildlife and their related parts and products",[1] has evolved into a critical international crisis. Despite dedicated efforts from officials,[2] non-profit organizations (NGOs),[3] and re-

[1]https://www.fws.gov/international/wildlife-trafficking/
[2]https://www.traffic.org/about-us/legal-wildlife-trade/
[3]https://www.eagle-enforcement.org/

"A superb 18th century European carved ivory dish. $11\frac{1}{2}$ in wide ... Estimate: £1500-2000 MENTION URL."

Figure 1: An example for wildlife product trading related post in the Online social networks. The subcaption is the text content where links and user mentions are masked. We investigate whether a post is WLT-related through its post text, images, and the linked webpages whenever necessary.

searchers, the illicit trade of endangered wildlife persists globally, thriving in lucrative black markets (Zimmerman 2003; Moyle 2009; Alacs and Georges 2008).

The chain of wildlife trafficking involves illegal capturing, organized transportation, and trading of wildlife and their products. This research concentrates on the final stage, wildlife trading, guided by the principle "No trading, No killing".[4] While the preceding stages primarily operate offline, wildlife trading has recently intertwined with the internet (Lavorgna 2014). The surge in online trading markets, especially through e-commerce websites, has exposed illegal wildlife trading to a wider audience worldwide (Sung and Fong 2018). Previous research has predominantly analyzed established online marketplaces like Etsy and Ebay (Sinovas et al. 2017; Pascual and Wingard 2021; Miller, Pay, and Smith 2019). In contrast, limited literature has explored the impact of wildlife trafficking on online social networks (Xu, Cai, and Mackey 2020; Xu et al. 2019). Moreover, we found the existing works analyzing wildlife trafficking in online social networks are preliminary, with

[4]https://tinyurl.com/yc786mh6

a lack of exploration into the abundant hidden information, including the multi-modal nature of selling posts. Systematically studying and leveraging these aspects to combat online crimes remain challenging. The pieces of evidence are also limited due to the scarce distribution nature of wildlife product sales posts. For example, Xu et al. (2019) filtered 138,357 suspicious tweets on Twitter, only finding 53 tweets from 38 unique users involved in ivory selling and zero pangolin-related posts. The limited number of accounts/tweets also makes building real-life applicable, effective machine learning less convincing. Under such a low recall rate, manually identifying wildlife product sales posts can be a Sisyphean effort. We urgently need a more efficient method to collect scalable data. The training and deployment of automatic learning algorithms that recognize potential illegal wildlife product trades can only become viable with such a method as a premise. Fig. 1 illustrates this challenge, with masked image and text content,[5] emphasizing the need for more efficient data collection methods.

To address these gaps, this research delves deeper into the patterns hidden behind selling posts on online social networks. For the first time, we present an Ivory-related multi-modal dataset, featuring positive samples and hard-negatives.[6] Leveraging a network-propagation-based method for data collection and a human-in-the-loop approach for labeling, our scalable and adaptive strategy can extend to other wildlife product categories with minimal effort. We benchmark machine learning results, introducing a practical framework that capitalizes on the multi-modality of the data. Furthermore, we offer rich observations and insights into distinguishing between the two classes, contributing to a more effective approach in identifying potential illegal wildlife product trades.

To this end, we make the following contributions:

- Data Scale Expansion: Beginning with minimal seed posts, we implement a network-propagation method to significantly expand the scale of our dataset, automating the collection of suspicious wildlife product trading (WLT) posts. This method is not only scalable but also highly adaptive, extending its applicability to various categories of WLT posts.

- Efficient Human-in-the-Loop Mechanism: We employ a human-in-the-loop mechanism to streamline the extraction of the most suspicious posts and their labeling. This approach efficiently reduces the laborious human effort required for identifying WLT posts and provides crucial hard negatives—normal/non-WLT posts that might be misclassified as WLT posts.

- First Scalable Dataset for OSNs[7]: We introduce and share a dataset focused on wildlife trading-related posts within online social networks (OSNs). To the best of our knowledge, this dataset is the first of its kind, offering scalability and empowering machine learning algorithms to automatically identify potential WLT posts—scarce yet profoundly impactful to the global ecosystem.

- Benchmarking Automatic Machine Learning Results: Using our dataset, we conduct a comprehensive benchmarking of automatic machine learning results, considering various modalities and design options. Our paper presents state-of-the-art results, supported by multiple evaluation metrics, advancing the field's understanding of WLT detection.

- Systematic Analysis of Patterns: We systematically analyze the distinctive patterns of WLT posts in comparison to hard-negative normal posts within online social networks. Our analysis provides valuable insights from multiple perspectives, serving as a foundation to attract the interest of researchers and paving the way for future works in this critical domain.

2 Related Work

Wildlife trafficking, acknowledged as the "Second-biggest direct threat to species after habitat destruction" by the World Wildlife Fund (WWF),[8] poses a severe global challenge. Over the years, illegal activities involving the capture and sale of endangered species have been meticulously documented. Wyatt (2021) extensively detailed the organized crime dynamics, profiling both offenders and victims. Legislators have also delved into the complexities of wildlife trafficking situations (Sollund 2019). Esteemed organizations like WWF,[9] TRAFFIC,[10] and EAGLE[11] actively combat violators engaged in the systematic illegal hunting, transportation, and sale of wildlife and their products (Dalberg 2012).

Historically, significant efforts have been invested in providing data (Gore et al. 2022, 2023b) and conducting analyses (Gore et al. 2019, 2023a) focusing on offline trafficking. However, with the advent of the internet, wildlife traffickers have adapted their strategies, actively promoting and selling products online (Lavorgna 2014). Most existing research and organizational reports concentrate on analyzing large online marketplaces (Sinovas et al. 2017; Pascual and Wingard 2021; Miller, Pay, and Smith 2019; Alfino and Roberts 2020; Cardoso et al. 2023). For instance, Miller, Pay, and Smith (2019) provide a systematic report on wildlife product trafficking across multiple online platforms. Alfino and Roberts (2020) present intriguing observations, noting that traffickers across European countries employ similar "code words" to facilitate cross-national trading.

While some studies explored wildlife product trading on the dark web, uncovering minimal evidence (Harrison, Roberts, and Hernandez-Castro 2016; Roberts and Hernandez-Castro 2017), only a handful have investigated promotions and sales on social networks (e.g., Twitter, Face-

---

[5]There are some examples in this paper for illustrative purposes in research. Authors did not intend to advertise the sales of wildlife products and tried their best to protect privacy.

[6]In this paper, we use "positive class" and "WLT class" interchangeably. Like wise, we use "hard-negatives", "negative class", and "normal class" interchangeably.

[7]https://github.com/GMouYes/WLT-OSN

[8]https://tinyurl.com/2p89zahk

[9]https://www.worldwildlife.org/

[10]https://www.traffic.org/

[11]https://www.eagle-enforcement.org/

book, Instagram, Pinterest). Wyatt et al. (2022) reported preliminary analysis on 500 private messages on Facebook and WhatsApp marketing 'exotic' pets collected from RENC-TAS.[12] However no data is publicly available and data collection methodologies remain unknown. Xu et al. (2019); Xu, Cai, and Mackey (2020) conducted preliminary research on Facebook and Twitter, identifying a mere 53 tweets promoting ivory-related products out of 138,357 collected. The low evidence recall rate (0.038%) highlights the inefficiency of manual verification in addressing wildlife promotion behaviors. Efficient methods surpassing simple keyword filtering mechanisms are urgently needed.

Our work differs from the prior work: 1) We devise a novel method for efficiently collecting and labeling a scalable dataset. The approach can be applied to other works with minimum adaptation effort. 2) We form and share the first dataset for WLT posts in OSNs, enabling possible automatic learning algorithms to be trained upon and applied to in-the-wild WLT posts. 3) We systematically analyze the unique characteristics of WLT posts in OSNs. We point out the unique multi-modality nature of the problem. Finally, we benchmark the dataset with multiple baselines and provide the current state-of-the-art design, which effectively identifies scarce however damaging, WLT posts.

3 Problem Formulation

Definition 1. WLT post: We define a wildlife product trading (WLT) related post on an online social network as a post that contains two key components:

- Discussion around wildlife products;
- Discussion around selling/buying these products.

For example, Fig. 1 explicitly shows a product made from ivory that is available for sale. On the other hand, only mentioning ivory (e.g., for education purposes) or selling other non-WLT-related products (e.g., wood) would be a normal post. We show more examples and discussions around the typical positive and negative posts in Sec. 7. In this work, we focus on ivory-related products and leave other types of wildlife products for future research.

Definition 2. WLT post identification: Given OSN posts X and label set $Y = \{0, 1\}$ with the following property

$$x = \{T, I, A\}, x \in X \tag{1}$$

where x contains the text T, images I, and other attributes A, such as user descriptions or other behavioral features. X is naturally multi-modal and multi-dimensional. The label with value 1 (positive class) represents WLT-related posts, while value 0 (negative class) represents normal posts. The WLT-related post-detection task aims to find the optimal model $f_\theta : x \to y$, which maps X to Y, and θ is the learnable parameters of the model.

4 Methodology

This section describes how we automatically collected suspicious data and adopted a human-in-the-loop method for labeling WLT posts and their hard-negative counterparts. We
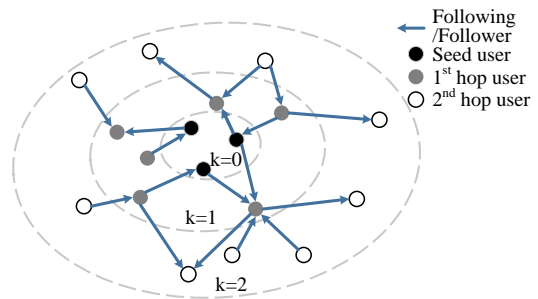
---

[12]https://renctas.org.br/home-en/

Figure 2: Illustration for Collecting Data. Nodes are the users, and edges represent their relationships. Given seed posts as nodes in black, we fetch their following/follower (blue and yellow edges) network for several hops. Eventually, we collect all these users' timelines as candidate data for further processing. Ideally, researchers can keep expanding the dataset scale by extracting more user hops, given their budget and computation limits.

note the following techniques are scalable (i.e., further expandable given the current data), adaptable (i.e., applicable to other wildlife products), and efficient compared to keyword filtering or naive human labeling methods.

4.1 Automate Data Collection

We present a comprehensive overview of our data collection process in Fig. 2. Commencing with nine seed tweets shared by Xu et al. (2019) pertaining to ivory-related Wildlife Product Trading (WLT) posts on Twitter, we acknowledge the significance of expanding the scale beyond these limited positive-only instances for the development of a practical and well-generalized machine learning application.

To address this limitation, we employ a network propagation approach to collect additional suspicious posts. Starting with the seed tweets, we retrieve the posting users, referred to as "seed users" and expand the network by collecting their followings, followers, and subsequent network layers. This scalable expansion process continues until the desired number of users is reached. Subsequently, we gather historical posts (up to 3,200 per user) from the collected users, resulting in a substantial dataset of suspicious posts for further selection and labeling.

The collection process is guided by two key insights derived from preliminary analysis:

- A user making one WLT post is likely to produce more WLT posts, justifying the search for additional posts from the same user.
- WLT sellers are more likely to be connected within the network, as individuals or groups of sellers may control multiple accounts involved in the same business.

Practically, we fetch the seed users and their two-hop neighbors based on the follower/following relationship and retrieve their posts, thus yielding an extensive dataset, comprising over *ten million* posts awaiting further processing.
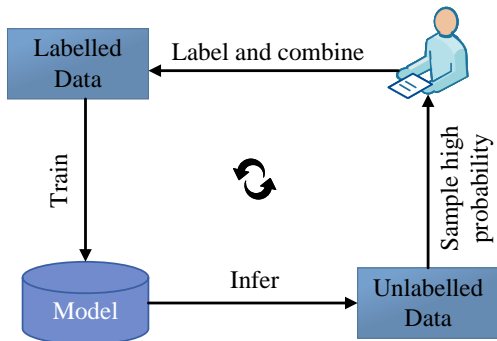
Figure 3: A human-in-the-loop process for labeling data.

Category	Features	Class		Overall
		Normal	WLT	
Linguistic	# Posts	8,421	255	8,676
Visual	# Posts w/ images	1,975	172	2,147
	# Posts w/ OCR	537	25	661
User	# Users posted	2,167	85	2,252
	# Users w/ profile	2,004	80	2,084
Interaction	# Posts w/ URLs	4,933	237	5,170
	# Posts w/ hashtags	2,740	149	2,889
	# Posts w/ mentions	5,100	75	5,175

Table 1: WLT Dataset Info.

Figure 4: Visualizing the deep learning framework for human-in-the-loop labeling process.

4.2 Human-in-the-loop Selection and Labeling

Although a substantial amount of data related to WLT is collected, the challenge lies in the large dataset size and a still-scarce positive rate, presenting an obstacle for human annotators. To overcome this, we implement an effective and efficient method for sampling and selecting highly probable WLT posts and labeling them.

Algorithm Design and Labeling The human-in-the-loop data selection and labeling mechanism, depicted in Fig. 3, involves the following steps:

1. Initial sampling of the most recent N posts from seed users, manually labeled, and combined with the seed posts to form the first labeled data group ($N = 100$).

2. Division of the labeled data into train/test sets for machine-learning model training.

3. Model inference on a large proportion of unlabeled data, assigning probability scores between 0 and 1 for each post's likelihood of being WLT.

4. Selection of the top K highest probability unlabeled data for another round of human labeling ($K \approx 2,500$).

5. Iterative repetition of steps 2-4 until the desired number of labeled posts ($\#posts \geq n$, where $n = 8K$).

6. Combining all labeled data, cleaning and filtering (with a focus on English-based content) the dataset.

During the labeling process, two of our authors served as annotators, bringing domain expertise and actively participating in discussions to define labeling criteria. Annotators independently labeled posts, with mutually agreed-upon labels being considered for adoption. Conflicting annotations were excluded from the final dataset. It's noteworthy that the human-in-the-loop data selection and labeling process is scalable. As more rounds are executed, machine learning models capture better signals for recognizing WLT posts, and human annotators label more highly suspicious posts. In our experiment, we conducted this process with 8K+ labeled data under the supervision of two human annotators.

Dataset Table 1 describes an overview of the collected dataset. We have 255 positive/WLT posts from 85 users and 8,421 hard negative/normal posts from 2,132 users.

It is worth mentioning that the dataset is naturally multi-dimensional and multi-modal, where each post contains text, images, and social-platform-specific interaction tokens (e.g., links, hashtags, and mentions of other users). Another obvious observation is that although we only include hard negatives as normal posts, the dataset is still highly unbalanced regarding almost all reported categories and features.

Model Design Our automatic learning algorithm is a specialized deep learning framework designed for the task, as depicted in Fig. 4. The framework comprises three major components: the *text encoder*, the *image encoder*, and the *decision maker*. Given the multi-modality nature of the task, the model takes both text contents and corresponding images as inputs, producing a score between 0 and 1—an indicative probability of the input post being a WLT post.

Text encoder: consumes the text content and generates its representation. We utilized pretrained language models such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) as our backbone for text encoding.

Image encoder: processes images and generates their corresponding representations. Our task, unique in handling varying numbers (0 to 4) of images per post, led us to two methods for addressing the multi-image problem: 1) Stitching: Down-sampling each image into 112×112 pixels and stitching the four images into one 224×224 image as input to the image encoder. 2) Concatenating: Using four images (without downsampling) as input to the image encoder, resulting in four distinct image representations. To achieve this, we employed pretrained vision models such as ResNet-

Modality	Model	Input	WLT		Overall		
			Pre.	Rec.	Macro F1	MCC	AUC
Single-Modal	Word Filter	Text	$.460_{.000}$	$\mathbf{1.000}_{.000}$	$.787_{.000}$	$.641_{.000}$	$.947_{.000}$
	BERT		$.724_{.027}$	$\underline{.942}_{.050}$	$.899_{.005}$	$.808_{.012}$	$\mathbf{.995}_{.003}$
	RoBERTa		$.813_{.029}$	$.835_{.031}$	$.903_{.016}$	$.807_{.032}$	$.981_{.005}$
	ResNet	$Image_S$	$.294_{.022}$	$.846_{.094}$	$.588_{.021}$	$.335_{.063}$	$.707_{.011}$
	ViT		$.268_{.051}$	$.692_{.196}$	$.569_{.047}$	$.248_{.134}$	$.696_{.020}$
	ResNet	$Image_C$	$.322_{.046}$	$.846_{.054}$	$.616_{.049}$	$.369_{.058}$	$.798_{.032}$
	ViT		$.316_{.040}$	$.862_{.148}$	$.614_{.041}$	$.373_{.101}$	$.786_{.024}$
Multi-Modal	BERT+ResNet	$Text + Image_C$	$\mathbf{.878}_{.135}$	$.855_{.090}$	$\underline{.922}_{.023}$	$\underline{.850}_{.045}$	$\mathbf{.995}_{.004}$
	BERT+ViT		$\underline{.821}_{.033}$	$.928_{.090}$	$\mathbf{.929}_{.026}$	$\mathbf{.860}_{.053}$	$\mathbf{.994}_{.002}$

Table 2: Experiment Results. We run each experiment three times and report the averages and the standard deviation (in underscript). In each column, best results are in bold and second best results are underlined.

50 (He et al. 2016) and Vision Transformers (ViT) (Dosovitskiy et al. 2020) as our backbone.

Decision maker: fuses text and image representations to derive the final output. Specifically, the decision maker concatenates all representations and employs a multi-layer perceptron with layer dropouts and ReLU activations. While recognizing alternative options, such as attention mechanisms, for vector fusions, we leave these variations for future exploration, focusing on paving the way for further technical advancements in this work.

All components' parameters are finetuned and updated altogether with loss backpropagation. We adopt a typical cross-entropy classification loss for training.

5 Benchmark Results

5.1 Models

We consider the following models as experiment baselines:

Single Modality Models: We first experiment with single-modality models, including text-only models that consume only text as inputs, and image-only models that consume only images as inputs.

Text-Only Models: We mainly used existing pretrained language models and finetune them on our dataset for downstream classification task. We also introduce a naive method serving as a lower-bound baseline due to its intuitively straightforward implementation.

- **Word Filter:** This naive method consumes the lowercased post text and only looks for the word "ivory". If the text contains such a word, the method will predict it as a WLT post. The naive approach is a lower bound for its straightforward design and low computation cost.
- **BERT:** We finetune a BERT (Devlin et al. 2019) on the dataset. While training large language models (LLMs) is too costly for many researchers, finetuning BERT is common and practical. More specifically, we finetuned the pretrained "BERT-base-uncased".
- **RoBERTa:** We also finetune RoBERTa (Liu et al. 2019) which claims to have a more robust performance in general. Specifically, we finetuned the pretrained "RoBERTa-base-uncased".

Image-Only Models: Similarly, we finetune two existing pretrained models in the computer vision (CV) domain, including convolutional-based methods and transformer-based methods. It is worth noting that the number of images per post can vary from zero to four, adding complexity to the research. For each tweet, we process the data in a padding-with-empty way, meaning that regardless of whether the tweet has attached images, we always create four placeholders for the images. We use all 0 to fill in the placeholders if there are insufficient images, resulting in a $3 \times 224 \times 224$ image for each tweet.

- **ResNet-50:** Convolutional neural networks (CNNs) have achieved remarkable success in image recognition tasks. We adopt the pretrained ResNet-50 (He et al. 2016), widely used due to its transferability.
- **ViT:** With the development of transformers in NLP domain (Vaswani et al. 2017), Vision Transformers (ViT) have emerged as a groundbreaking architecture (Dosovitskiy et al. 2020). ViTs break down an image into fixed-size patches and linearly embed each patch. We applied the pretrained "ViT-base" for finetuning.

Bi-Modality Models: We experiment with variations of our design illustrated in Sec. 4.2 for bi-modality as below:

- **Image processing variants:** $image_S$ stands for image stitching method while $image_C$ accounts for image flattening/concatenation method.
- **Text processing variants:** We had pure text input, text+OCR (i.e., post text and text extracted from images) results as text input, and text+OCR+user profile description as text input variants.

5.2 Experiment Settings

To address the dataset imbalance, we down-sample the negative class to keep the class ratio between WLT and normal posts as 1:10 (i.e., We sample 2,550 normal posts from the complete set of negative class). We extract all posts containing "ivory", then randomly sample the rest of the instances to meet the requirement. The data is split into train/dev/test sets (70/20/10 ratio), with the constraint that posts from the same user can only show up in at most one split. We then combine the corresponding splits of both classes to form the final data split. The intuition behind such a splitting scheme

Modality	Model	Input	WLT		Overall		
			Pre.	Rec.	Macro F1	MCC	AUC
Multi-Modal-$Image_C$	BERT+ResNet	Text + Images	$.694_{.192}$	$.814_{.123}$	$.846_{.045}$	$.713_{.074}$	$.975_{.007}$
		Text + Images + OCR	$\mathbf{.878}_{.135}$	$.855_{.090}$	$\underline{.922}_{.023}$	$\underline{.850}_{.045}$	$\mathbf{.995}_{.004}$
		Text + Images + OCR + Desc.	$.816_{.058}$	$.754_{.050}$	$.882_{.017}$	$.765_{.033}$	$.982_{.002}$
	BERT+ViT	Text + Images	$.810_{.165}$	$.643_{.106}$	$.824_{.024}$	$.678_{.034}$	$.977_{.008}$
		Text + Images + OCR	$.821_{.033}$	$\underline{.928}_{.090}$	$\mathbf{.929}_{.026}$	$\mathbf{.860}_{.053}$	$\mathbf{.994}_{.002}$
		Text + Images + OCR + Desc.	$.686_{.206}$	$.725_{.176}$	$.819_{.004}$	$.659_{.015}$	$.973_{.014}$
Multi-Modal-$Image_S$	BERT+ResNet	Text + Images	$.782_{.050}$	$.783_{.075}$	$.881_{.034}$	$.763_{.067}$	$.982_{.005}$
		Text + Images + OCR	$.763_{.148}$	$.812_{.181}$	$.871_{.011}$	$.756_{.021}$	$.991_{.002}$
		Text + Images + OCR + Desc.	$.668_{.068}$	$\mathbf{.957}_{.075}$	$.879_{.015}$	$.777_{.019}$	$.990_{.009}$
	BERT+ViT	Text + Images	$.803_{.130}$	$.806_{.035}$	$.886_{.051}$	$.780_{.093}$	$.983_{.013}$
		Text + Images + OCR	$\underline{.854}_{.052}$	$.872_{.110}$	$.921_{.040}$	$.847_{.073}$	$.993_{.005}$
		Text + Images + OCR + Desc.	$.732_{.257}$	$.725_{.153}$	$.830_{.036}$	$.682_{.064}$	$.980_{.011}$

Table 3: Detailed Experiment Results.

is: 1) within each class, we prevent user information leakage. 2) On the contrary, we allow posts from the same user in different classes to show up in different splits, as a user can post WLT as well as normal ones. We train/finetune models on the training set using various hyperparameter combinations, selecting the best-performing model on the dev set for evaluation on the test set. All models are run on machines with Nvidia GeForce 3090 GPUs.

5.3 Evaluation Metrics

We evaluate model performance with these metrics:

- For the concerned positive class (WLT), we measure the precision (Pre.) and recall (Rec.) rate.
- For the overall evaluation, we measure the Macro F1, Matthews Correlation Coefficient (MCC), and Area Under the Curve score (AUC).

We run each model with best-performing hyperparameters under three random seeds and report their metric results with average scores and standard deviation. As the data is still imbalanced, we mitigated the issue with inverse class weights, and then we recalibrated the classification threshold towards maximizing the MCC score on the validation set.

5.4 Main Results

Table 2 presents the results of our experiments, and noteworthy observations are highlighted below:

Naive method has shortcomings despite some merit: The naive method (word filtering) achieved a perfect positive class recall rate and a relatively high AUC. However, its limitation lies in its inability to recognize positive posts lacking the keyword, leading to misclassifications of negative posts containing the keyword. This results in lower Macro F1 and MCC scores, emphasizing the constraints of word filtering without considering post context. It's essential to note that the "Word Filter" method is deterministic, resulting in results with a standard deviation of 0.

Text modality has an edge over vision: In the comparison between text-based and image-based single-modal models, text-based models exhibited superior positive class precision and overall performance. This advantage is partially attributed to many samples lacking images in the original posts (see Fig.9 and Sec.6.3 for details). Despite this, visual models demonstrated merit, especially in multi-modal settings, as discussed in the following paragraph. Additionally, ViT consistently outperformed ResNet under both image concatenation and image stitching settings. Furthermore, each model generally performed better with image concatenation than image stitching.

Multi-modality models achieve overall best results: Moreover, we list the best-performing results for our bi-modal variations. For the bi-modality models, we observe the best precision result for the concerned positive class (0.878 for BERT+ResNet and 0.821 for BERT+ViT). Meanwhile, we also observe the best performance for BERT+ViT across all three overall metrics and BERT+ResNet achieves similar/second best performance. The best multi-modal result (BERT+ViT) gained an improvement against the best single-modal results of 2.9% on Macro F1 and 6.4% on MCC. Such a strong result indicates embracing the advantage from both modalities effectively improves the performance under various metrics.

5.5 Multi-modal Variation Results

We conducted additional experiments exploring various modeling possibilities for multi-modal frameworks, including image OCR (Optical Character Recognition), user descriptions, and their combinations with the previously mentioned model variations in Table 3. Key findings include:

- OCR proved beneficial in extracting relevant information embedded in images, such as product descriptions, sales prices, and selling websites. This enhancement contributed to overall improved performance.
- The inclusion of user descriptions had a negative impact on model performance, potentially due to the limited number of users. With fewer posting users than posts, the model tended to overfit on user attributes.
- Similar to the single modality results, multi-modal results w/ ViT consistently outperform those w/ ResNet.

These benchmarking results and insights lay the foundation for further improvements in classification results, which can be explored in future work.

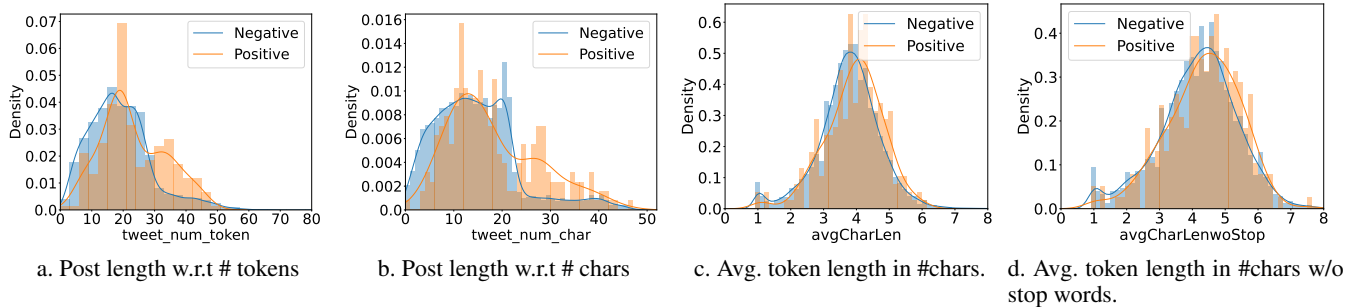| a. Post length w.r.t # tokens | b. Post length w.r.t # chars | c. Avg. token length in #chars. | d. Avg. token length in #chars w/o stop words. |

Figure 5: Distribution of text length for (a) positive and (b) negative posts, as well as average token length (c) w/ and (d) w/o stop words. It is worth noting that normal posts tend to have long tails on the post length and average token length, but we had cutoffs in the visualizations to focus on the left parts. For more information on the max length, please refer to Table 4.

6 Dataset Analysis

This section unveils intriguing insights gleaned from the labeled dataset, as showcased in Table 1. We present our findings from the analysis of textual and visual modalities in Sec. 6.1 and 6.3, respectively. Additionally, we delve into the distributions of OSN-specific tokens, such as mentions, URLs, and hashtags, in Sec. 6.2. While our preliminary experiments primarily focused on harnessing deep learning methods for classification, these observations serve as valuable guidance for future research endeavors, particularly in the development of essential handcrafted features for general machine learning frameworks.

| Category	WLT			Normal		
	avg.	std.	max	avg.	std.	max
#words	23	10.3	49	19	10	78
#words w/o ST	18	10	47	16	10	78
#chars	91	48	232	72	43	246
#chars w/o ST	86	48	231	69	43	244
#char/word	3.9	0.9	6.1	3.8	1.0	18
#char/non-ST	5.1	1.1	11	4.9	1.6	18
#char/non-SW	4.3	1.1	7.5	4.2	1.2	18
#char/non-SW/ST	5.9	1.3	11	5.7	1.9	18

Table 4: Dataset Text Statistics. ST: special tokens (retweet/rt, mentions, hashtags, urls, etc.). SW: stop word.

6.1 Text Analysis

We present comprehensive text analysis statistics in Table 4. Additionally, detailed distributions of text features within three major categories: writing styles, text quality, and text semantics, are illustrated in finer granularity through figures in the following subsections.

Writing Style WLT posts differ from normal posts in their intention, reflecting the variance in writing styles. We analyze the writing style differences concerning post length and average word length. NLTK (Loper and Bird 2002) Tweet tokenizer is employed for tokenizing text into words.

Post Length: The distribution of text posts is presented in Table 4 and Fig. 5 from two perspectives: post length in terms of the number of tokens and characters. Generally,
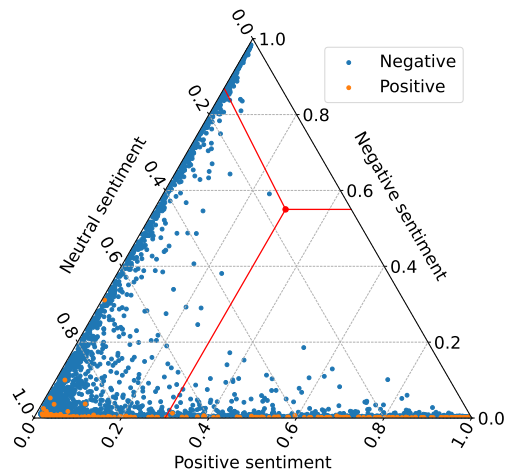
a. Sentiment Analysis for posts.

b. Toxicity Analysis for posts.

Figure 6: Semantic Analysis for posts.

positive posts are longer on average, typically displaying two local maxima in distributions. Negative samples, while having a lower average post length, exhibit a long tail accounting for more extreme cases. This difference results in a wider range of text lengths for negative samples.

Word Length: Average token/word length in the number of characters per post is considered. The results are shown in Table 4. Positive posts tend to use longer tokens, but negative samples have a long tail for extreme cases. The impact of stop words and unique tokens on OSNs is showcased, with minimal influence on the relative comparison results. Distribution visualizations in Fig. 5 support these findings.

Figure 7: Word cloud for pos.(top) & neg.(bottom) posts.

Text Semantics We investigate the nature of text semantic differences between the two classes, focusing on sentiments and toxicity levels in Fig. 6a and Fig. 6b.

Sentiment Analysis: VADER (Hutto and Gilbert 2014) is used for evaluating text sentiments. Each dot in Fig. 6a represents a post with three-dimensional scores (positive, neutral, and negative sentiment score). WLT posts predominantly exhibit positive/neutral sentiments, while normal posts display a more even distribution of all sentiments.

Toxicity Analysis: The pretrained deep learning model DeToxifying (Bose, Perera, and Dorr 2023) measures text toxicity levels. Average toxicity scores of WLT posts and normal posts are compared in Fig. 6b. Both posts mostly have relatively low toxicity across all categories, with WLT posts showing even lower toxicity scores due to their product-promoting intentions.

Frequent Word Usage We conduct an analysis of the most commonly used words in both classes, illustrating word clouds of WLT and normal posts in Fig. 7, where larger fonts signify higher occurrence within the dataset. The presence of shared frequent words such as "ivory" underscores the constraints of simplistic keyword filtering methods. Further validation in the experiment and case study sections yields both quantitative and qualitative results, enriching our understanding of the efficacy of our approach.
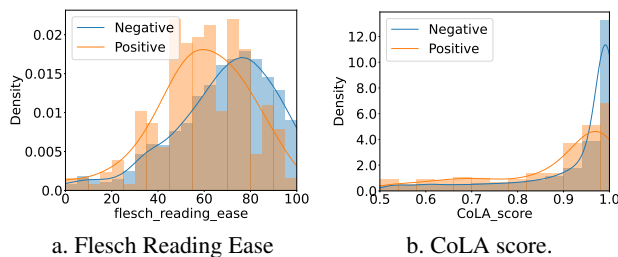

a. Flesch Reading Ease   b. CoLA score.

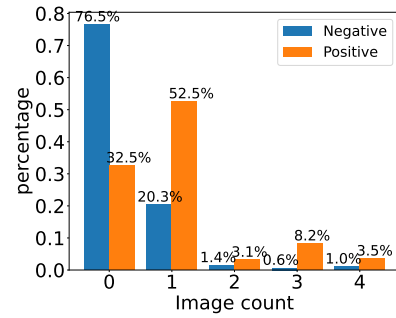Figure 8: Text Readability and Linguistic Acceptability.

Figure 9: Distribution of # images in posts.

Text Quality Another important perspective for the posts is their text quality. We analyze the posts' language quality in two ways: the text readability and the linguistic acceptability. More specifically,

- We measure text readability with Flesch Reading Ease Score (Flesch 1979), computed according to Eq. 2.

$$s = 206.835 - 1.015(\frac{words}{sentences}) - 84.6(\frac{syllables}{words}) \quad (2)$$

  The Flesch Reading Ease typically ranges between 0 and 100, where a higher score means the text is easier to read.

- We leverage a language model BERT (Devlin et al. 2019) finetuned on the CoLA (Warstadt, Singh, and Bowman 2018) dataset, specifically focused on predicting language acceptability.[13] The finetuned model will infer on our data, giving each sample a score between 0 and 1, where a higher score indicates higher linguistic acceptability. 0 means not acceptable, while 1 means perfectly acceptable text in the linguistic sense.

The distribution results are shown in Fig. 8. We noticed that: 1) the majority of the posts, regardless of their labels, are within decent scores for both readability and linguistic acceptability; 2) The positive posts tend to have lower readability scores and linguistic acceptability scores; These observations may be because positive posts are more prolonged and use more complicated words (thus less readability and linguistic acceptability intuitively).

Category	WLT			Normal		
	avg.	std.	max	avg.	std.	max
# URLs	1.5	0.6	3	0.7	0.7	4
# mentions	0.3	0.5	2	0.96	1.7	50
# hashtags	3.3	3.9	23	1.0	2.2	24

Table 5: Dataset Special Token (ST) Statistics.

6.2 Interactions and Special Activities

Special tokens' statistics are provided in Table 5, with detailed distributions presented in Fig. 10. Special tokens include links, mentions, and domain-specific hashtags. WLT

---

[13]https://tinyurl.com/tcdnj59a

a. # Hashtag Distribution   b. # Mention Distribution   c. # URL Distribution
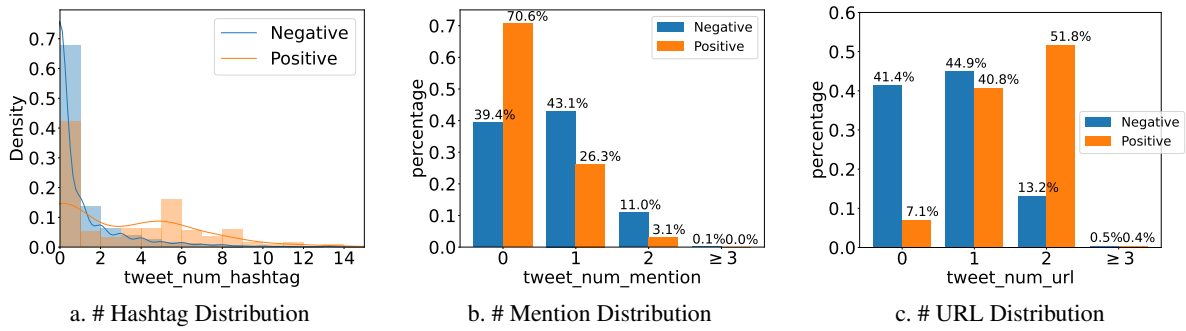
Figure 10: Distribution of special tokens. Note that we truncate Fig. 10a positive class tails for better visualization.

posts tend to have more links and hashtags for product promotion purposes but fewer mentions and reposts, aligning with the expectation of broadcasting to a wider audience rather than specific user notifications.

6.3 Image Analysis

In the following subsections, we perform image analysis on images in WLT (positive) posts and normal (negative) posts.

Image Number Statistics Our selected $8,676$ tweets contain $2,713$ images and $2,147$ tweets have at least one image as shown in Table 1 . Detailed distribution is shown in Fig. 9, showing that 76.5% normal posts (negative class) do not contain images. On the other hand, WLT posts (positive class) consistently have more images (1, 2, 3, and 4 images).
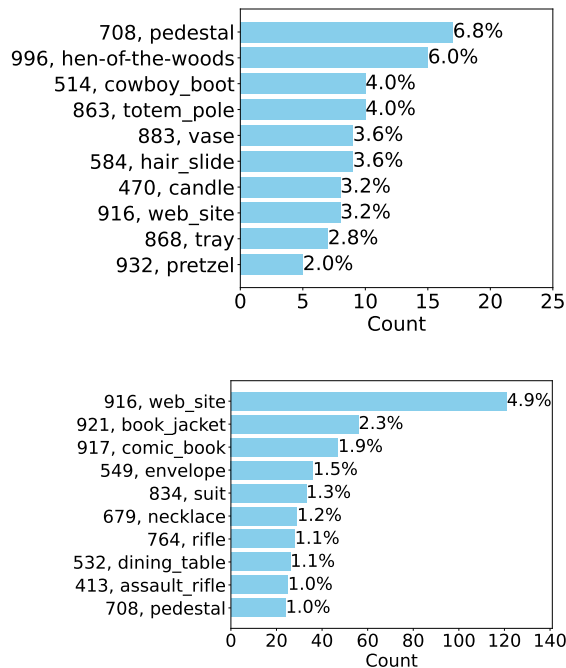

Figure 11: Classification result for pos.(up) & neg.(bottom).

Classification We employed ResNet-50 (He et al. 2016), pretrained on IMAGENET1K_V1 (Krizhevsky, Sutskever,

and Hinton 2012), to classify images. Fig.11 displays the top 10 classes classified by ResNet-50 for positive and negative posts' images. Positives exhibit a distinct class distribution compared to negatives. The prevalent images in positives are closely associated with the shape, material, or color of ivory carving products. In contrast, the most common images for the entire dataset and negatives are related to the "web_site" class, as expected, typically containing events-related or product promotion-related information. Additionally, there are some recurring classes related to books.

7 Case Study

In this section, we present non-trivial examples to underscore the complexity of the wildlife product trading problem. Fig. 12 showcases various instances, with the first row illustrating different Wildlife Product Trading (WLT) posts and the second row depicting typical normal posts.

7.1 Positive Cases

In addition to the WLT post shown at the beginning of the paper in Fig. 1, we explore further examples down below.

WLT without "ivory" in text: Fig.12a exemplifies that a WLT post may not necessarily contain the term "ivory". This challenges naive word filtering methods and models biased toward the term "ivory" in the positive class, potentially leading to false negatives. Traders may deliberately avoid explicit terms, opting for alternative "code words" to evade detection(Alfino and Roberts 2020).

WLT but not elephant ivory: Ivory is not exclusive to elephants; it can originate from various animals. Fig.12b and Fig.12c showcase examples of WLT posts involving products from animals such as walruses and rhinos. This highlights the involvement of WLT product sellers in multiple illegal sales activities.

WLT with sales information embedded in images: Fig. 12d demonstrates cases where extensive information is embedded in images, and the post class can be inferred based on images alone. This information often includes product type, price, and selling links, providing essential elements for identifying WLT-related posts.

7.2 Negative Cases

Our dataset includes many non-trivial hard-negative cases, presenting challenges in identification:

a. "March Estate Art, Antique ... TIMED ONLINE Auction - URL URL"

b. "... Ivory and Rhinoceros antiquities ... sold, paid and collected ... Despite the potential ban, the market is still strong ... URL"

c. "... 3 wonderfully detailed ... ancient walrus ivory. For sale on website. URL"

d. "Mammoth Ivory Carvings Figurine of Japanese Samurai ... precisely sculpted with original mammoth ivory. URL URL

e. "HB5578 will require documentation the ivory ... hoping that will save elephants in Africa??!! URL"

f. "... This Antiqued French Candelabra is ... in a beautiful patina of grey, gold and ivory. URL URL"

g. "Sumptuous Ivory Silk Quilt Bedspread URL URL"

h. "Special thank you to ... followers from Ivory Coast ... URL URL"

Figure 12: WLT posts (a. – d.) and normal posts (e. – h.) for case study.

Displaying ivory products but not for trades: Fig. 12e shows an account engaging in a discussion on whether ivory products should be banned. This post serves a different purpose than trading and is not labeled as a WLT post.

Ivory as a color: Ivory can refer to a color close to white, causing confusion for word filtering methods. Fig.12f and Fig.12g exemplify instances where the term "ivory" post challenge in distinguishing contexts.

Special Phrases: Fig. 12h includes the term "Ivory Coast", the official name of the Republic of Côte d'Ivoire. Special phrases like this can lead to false positives for models biased on "ivory" without capturing the post's actual context.

Other Cases: During our research, we encountered posts related to products like "rosewood"(the ivory of forest), often associated with concerns and protection. While these are not considered positive cases in our research unless occurring simultaneously with ivory products, they highlight additional challenges in classifying wildlife product-related posts.

8 Limitation

Our research is constrained by several limitations. Our collected dataset mainly focuses on a specific type of wildlife product: ivory-related. While there may be some differences between other types of wildlife products, our data collection, filtering, labeling, and classification modeling methods are primarily adaptable to recognizing other wildlife products. Our data collection method is network-propagation-based. Thus, the characteristics of the other isolated network clusters may not be best captured. However, given any other

clusters' seed posts, we can extend the scope and establish a similar dataset given our designated approach. Lastly, it is hard to identify the actual trading record, as the real trade is accomplished on the redirected e-commerce websites or offline sites. We thus define our identification scope as the advertising for potential or finished sales and the malicious intention of selling/buying these products.

9 Conclusion

Our investigation delves into wildlife product trading behaviors on online social networks, pioneering novel methodologies for data collection, filtering, and labeling. This endeavor culminates in the creation of the inaugural dataset focusing on ivory-product trades within the online social network sphere. Through meticulous analysis of the proposed dataset, we rigorously benchmark machine learning results based on the collected data. Notably, our approach exhibits scalability and adaptability to other wildlife product trade identification tasks with minimal additional effort. For future endeavors, we envision collaborations with domain experts to develop more expansive datasets, leveraging information retrieval technologies and deep learning methodologies. Additionally, we aim to design enhanced algorithms based on the insights gleaned from the proposed dataset. While our research concentrates on WLT behaviors in online social networks, we note its potential in adapting to other criminal marketplaces. We leave these avenues for exploration in future research endeavors.

References

Alacs, E.; and Georges, A. 2008. Wildlife across our borders: a review of the illegal trade in Australia. *Australian Journal of Forensic Sciences*, 40(2): 147–160.

Alfino, S.; and Roberts, D. L. 2020. Code word usage in the online ivory trade across four European Union member states. *Oryx*, 54(4): 494–498.

Bose, R.; Perera, I.; and Dorr, B. 2023. Detoxifying Online Discourse: A Guided Response Generation Approach for Reducing Toxicity in User-Generated Text. In *SICon*.

Cardoso, A. S.; Bryukhova, S.; Renna, F.; Reino, L.; Xu, C.; Xiao, Z.; Correia, R.; Di Minin, E.; Ribeiro, J.; and Vaz, A. S. 2023. Detecting wildlife trafficking in images from online platforms: A test case using deep learning with pangolin images. *Biological Conservation*, 279: 109905.

Dalberg, W. 2012. Fighting illicit wildlife trafficking: A consultation with governments. *WWF, Gland, Switzerland*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *NAACL-HLT*, 4171–4186. Association for Computational Linguistics.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Flesch, R. 1979. *How to write plain English: A book for lawyers and consumers*, volume 76026225. Harper & Row New York.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gore, M. L.; Braszak, P.; Brown, J.; Cassey, P.; Duffy, R.; Fisher, J.; Graham, J.; Justo-Hanani, R.; Kirkwood, A. E.; Lunstrum, E.; et al. 2019. Transnational environmental crime threatens sustainable development. *Nature Sustainability*, 2(9): 784–786.

Gore, M. L.; Griffin, E.; Dilkina, B.; Ferber, A.; Griffis, S. E.; Keskin, B. B.; and Macdonald, J. 2023a. Advancing interdisciplinary science for disrupting wildlife trafficking networks. *PNAS*, 120(10): e2208268120.

Gore, M. L.; Hilend, R.; Prell, J. O.; Griffin, E.; Macdonald, J. R.; Keskin, B. B.; Ferber, A.; and Dilkina, B. 2023b. A data directory to facilitate investigations on worldwide wildlife trafficking. *Big Earth Data*, 7(2): 338–348.

Gore, M. L.; Schwartz, L. R.; Amponsah-Mensah, K.; Barbee, E.; Canney, S.; Carbo-Penche, M.; Cronin, D.; Hilend, R.; Laituri, M.; Luna, D.; et al. 2022. Voluntary consensus based geospatial data standards for the global illegal trade in wild fauna and flora. *Scientific Data*, 9(1): 267.

Harrison, J. R.; Roberts, D. L.; and Hernandez-Castro, J. 2016. Assessing the extent and nature of wildlife trade on the dark web. *Conservation Biology*, 30(4): 900–904.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*, volume 8, 216–225.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25.

Lavorgna, A. 2014. Wildlife trafficking in the Internet age. *Crime Science*, 3(1): 1–12.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loper, E.; and Bird, S. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Miller, C.; Pay, J.; and Smith, J. 2019. Detecting online environmental crime markets. Digital dangers report.

Moyle, B. 2009. The black market in China for tiger products. *Global Crime*, 10(1-2): 124–143.

Pascual, M.; and Wingard, J. 2021. ASSESSING ONLINE WILDLIFE MARKETING.

Roberts, D. L.; and Hernandez-Castro, J. 2017. Bycatch and illegal wildlife trade on the dark web. *Oryx*, 51(3): 393–394.

Sinovas, P.; Price, B.; King, E.; Hinsley, A.; and Pavitt, A. 2017. Wildlife trade in the Amazon countries: an analysis of trade in CITES listed species. *Technical report prepared for the Amazon regional program. UN Environment, Cambridge*, 112.

Sollund, R. A. 2019. *The crimes of wildlife trafficking: Issues of justice, legality and morality*. Routledge.

Sung, Y.-H.; and Fong, J. J. 2018. Assessing consumer trends and illegal activity by monitoring the online wildlife trade. *Biological Conservation*, 227: 219–225.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Warstadt, A.; Singh, A.; and Bowman, S. R. 2018. Neural Network Acceptability Judgments. *arXiv preprint arXiv:1805.12471*.

Wyatt, T. 2021. *Wildlife trafficking: a deconstruction of the crime, victims and offenders*. Springer.

Wyatt, T.; Miralles, O.; Massé, F.; Lima, R.; da Costa, T. V.; and Giovanini, D. 2022. Wildlife trafficking via social media in Brazil. *Biological Conservation*, 265: 109420.

Xu, Q.; Cai, M.; and Mackey, T. K. 2020. The illegal wildlife digital market: an analysis of Chinese wildlife marketing and sale on Facebook. *Environmental Conservation*, 47(3): 206–212.

Xu, Q.; Li, J.; Cai, M.; and Mackey, T. K. 2019. Use of machine learning to detect wildlife product promotion and sales on Twitter. *Frontiers in big Data*, 2: 28.

Zimmerman, M. E. 2003. The black market for wildlife: combating transnational organized crime in the illegal wildlife trade. *Vand. J. Transnat'l L.*, 36: 1657.

A Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. Throughout the research process, we respect user privacy and collected public data from online social networks. We tried to fight against wildlife trafficking, which is widely prohibited across majority of regions.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.

   (e) Did you describe the limitations of your work? Yes.

   (f) Did you discuss any potential negative societal impacts of your work? No. We did not find any potential negative societal impacts of this work.

   (g) Did you discuss any potential misuse of your work? No. We did not find any potential misuse of our work.

   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes.

   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

   (a) Did you clearly state the assumptions underlying all theoretical results? NA.

   (b) Have you provided justifications for all theoretical results? NA.

   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA.

   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA.

   (e) Did you address potential biases or limitations in your theoretical framework? NA.

   (f) Have you related your theoretical results to the existing literature in social science? NA.

   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA.

3. Additionally, if you are including theoretical proofs...

   (a) Did you state the full set of assumptions of all theoretical results? NA.

   (b) Did you include complete proofs of all theoretical results? NA.

4. Additionally, if you ran machine learning experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes.

   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.

   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes. We provided several metrics measuring and reflecting the mis-classify rates.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? Yes. We tested several baselines as existing models. We leveraged pretrained image and text encoders in our framework. All of the above-mentioned models are cited.

   (b) Did you mention the license of the assets? NA.

   (c) Did you include any new assets in the supplemental material or as a URL? No.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes. All data we retrived are from online social networks, with public API access for research purpose. All models we leveraged are allowed for research purpose.

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. During our research, we masked personally identifiable information to respect privacy. We did not find strong evidence for offensive content, as is partially shown in our sentiment analysis section.

   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR)? Yes.

   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes. More details are shown in the data repository.

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots? NA

   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

   (d) Did you discuss how data is stored, shared, and deidentified? NA

B Image Classification Examples

C Posting Users

We provide some straightforward analysis on the posting users in Fig. 15. We define the "WLT users" as those had at least one WLT post, and the others as "normal users". Note that WLT users can also post normal posts. We plot the inter-connections among WLT users in Fig. 15a, the inter-connections among normal users in Fig. 15b, and the intra-connections between WLT users and normal users in Fig. 15c. The following and follower connections are shown in two different edge colors. To specify the differences, we plot the in-degree and out-degree distributions separately. Note that we downsampled the normal users to 100 such

Figure 13: Top-10 image class examples in the positive/WLT class.

Positive Class	Count	Negative Class	Count
pedestal	17	web_site	121
henofthewoods	15	book_jacket	56
cowboy_boot	10	comic_book	47
totem_pole	10	envelope	36
vase	9	suit	33
hair_slide	9	necklace	29
candle	8	rifle	28
web_site	8	dining_table	26
tray	7	assault_rifle	25
pretzel	5	pedestal	24

Table 6: Top-K Image Classification Result.

ysis for future work. Although the analysis opens a new discussion opportunity around user-level wildlife product trading identification task, this paper mainly focused on post-level identification problem.
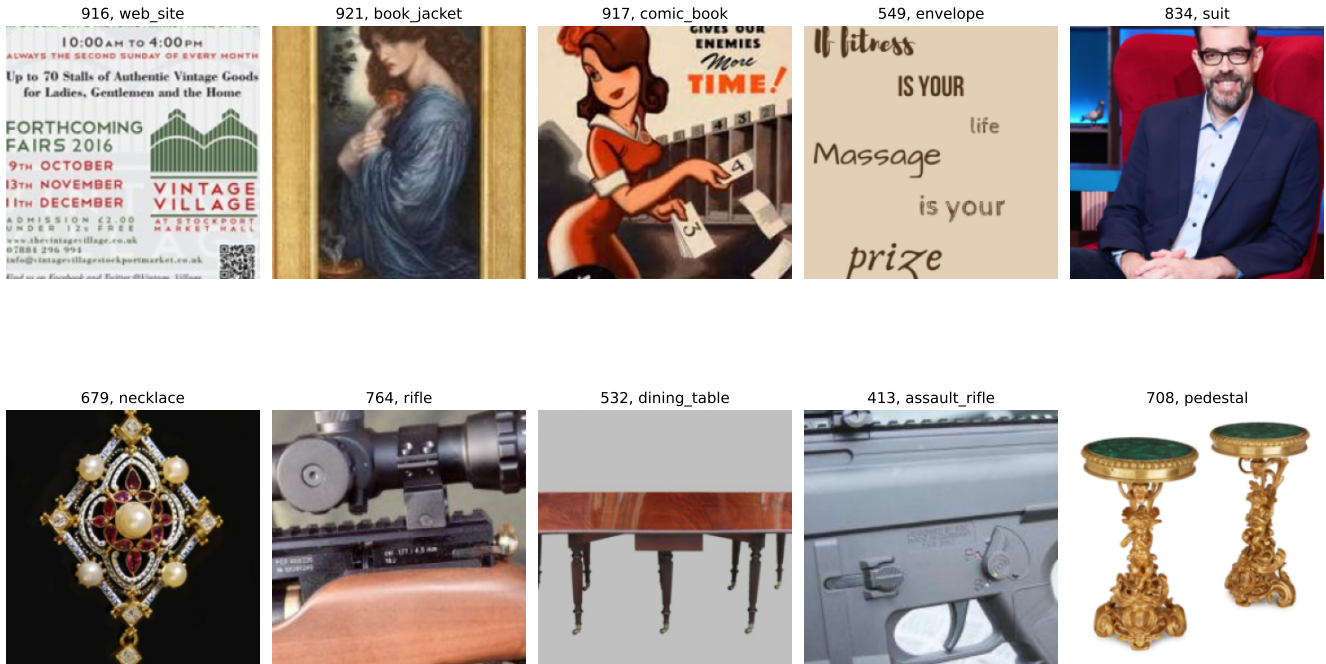
that they have the similar scale in # users as WLT users. We observe that WLT users tend to have higher in-degree and out-degree extreme cases (the left-most parts in rank plots). However, in a more general scope, the WLT users tend to have less connections than the normal users.

We are aware of other user definitions (e.g., generalize at least one WLT-post to at least k WLT-posts) or other user analysis perspectives (e.g., username, user profile description, user verified status, user profile image analysis, etc.). We focused on the network analysis and leave the other anal-

Figure 14: Top-10 image class examples in negative/normal class.

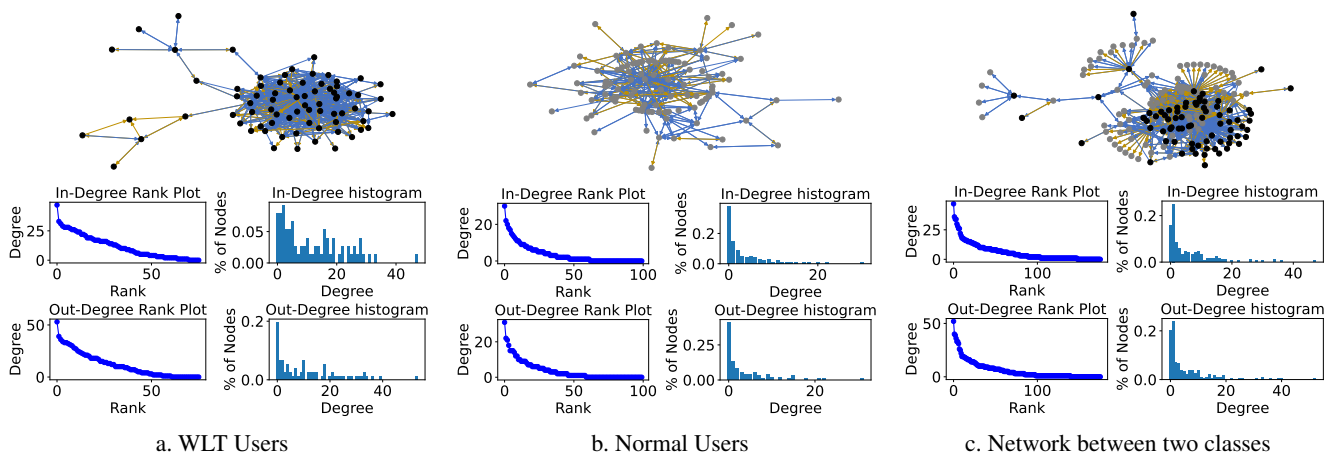a. WLT Users   b. Normal Users   c. Network between two classes

Figure 15: Connections and distributions of users.