Lost in Translation: Analyzing Non-English Cybercrime Forums

Mariella Mischinger[1,2], Jack Hughes[3], Fedor Vitiugin[4], Sergio Pastrana[2], Alice Hutchings[3], Guillermo Suarez-Tangil[1]
[1]IMDEA Networks Institute, Spain
[2]Universidad Carlos III de Madrid, Spain
[3]University of Cambridge, United Kingdom
[4]University of Turku, Finland

***Abstract*—Cybercrime analysis and Cyber Threat Intelligence are crucial for understanding and defending against cyber threats, with online underground communities serving as a key source of information. Classification tasks are popular but demand significant manual effort and language-specific expertise. Prior work focuses on English-language forums, as non-English languages require fluent domain experts. We evaluate machine translation tools for suitability in preserving contextual information in posts and find GPT-4 is most reliable. We leverage existing underground forum post classification pipelines to compare their performance on translated text and original language text. We find classification performed on translated underground forum data is as effective as on original language text, enabling researchers to reuse existing pipelines. Finally, we investigate a fully machine-generated few-shot and zero-shot classification to reduce reliance on manual labeling, followed by a two-step machine-based classification, combining machine-generated labels with the existing classification pipeline. We find machine-based labeling causes errors to propagate downstream. For tasks requiring high-quality label creation, human expertise remains essential. Finally, we provide a qualitative evaluation of disagreements in annotator labels of the original language and the translations, as well as disagreements between annotators and machine labeling.**

I. Introduction

Cyber Threat Intelligence (CTI) plays a critical role in preparing for and defending against cyberattacks. Its effectiveness relies on a proper understanding of modern threats. An emerging source of intelligence are online underground communities, where offenders gather and share knowledge, exchange tools, and discuss novel attack methods [1], [2], [3]. Monitoring forums allows researchers to understand emerging cybercrimes [4], [5], and acquire valuable information on the initial stages of cyberattacks, i.e., planning and development [6]. While these forums can be accessed worldwide, they tend to specialize in dedicated communities, using one or more distinct languages [7], [1], [8].

Due to the vast amount of data in these forums, researchers often rely on automated classification and evaluation approaches [9], [10], [11], [12], [13]. Such approaches use the linguistic characteristics of forum posts [14], [15]. However, most of the existing classification approaches focus on English and are not suitable for multilingual data [14], [15], [5], [6]. This creates an important gap as language-based segmentation can obscure region-specific cybercrime activities, limiting the effectiveness of detection methods that account for other languages [16], [17]. For example, some Russian-speaking regions are renowned cybercrime hubs. As such, many prominent underground hacking forums operate in Russian and serve as key platforms for establishing "partnerkas."

To gather intelligence from multilingual communities, it is essential to design and evaluate efficient approaches that break existing language barriers. Machine translation (MT) is currently the de facto mechanism [6], [18]. However, the language used in these forums is highly specialized and context-dependent [19], [20], [21], raising questions about whether loose translations hinder accurate interpretation, either by human analysts or automated systems processing translated content [22], [23]. Advances in NLP have introduced new possibilities for text classification, particularly through the development of multilingual language embeddings and large language models (LLMs). However, their effectiveness in interpreting cybercriminal discourse remains uncertain, especially given known limitations in accurately identifying fundamental stances [24]. These constraints are further amplified with low-resource or less widely spoken languages, where qualified annotators are scarce [7] or entirely unavailable [14], [25], [5], [26], [8], [27]. As such, our work is motivated by a central question: *Given the difficulties of finding human annotators for certain languages, what alternative approaches can be employed to obtain annotations of comparable quality?*

In this work, we investigate various strategies for systematically processing *multilingual* data from underground forums. We begin by assessing different MT systems to determine the most suitable option for forum content, based on assessments from domain experts. Subsequently, we adopt established labeling frameworks for underground forums and examine two processing pipelines: (1) using multilingual or language-specific NLP models to analyze the original language; and (2) translating the content into English for processing with English-centric NLP models. Finally, we explore alternatives to manual annotation by assessing automated labeling methods, including: (3) fully machine-driven zero-shot and few-shot classification using LLMs, and (4) a *2-step machine based classification* approach, where LLM-generated pseudo-labels

979-8-3315-8969-1/25/$31.00 ©2025 IEEE

are used to train a conventional supervised classifier.

By investigating these strategies using a comparative methodology (§III) over a dataset of underground forums labeled by native speakers of multiple languages (§IV), we make the following contributions:

- We investigate the performance of four translators, including MT models and LLMs (§V). Our dataset spans five languages: Arabic, German, Russian, Spanish, and Vietnamese. We recruit domain experts who are proficient in English and are native speakers of one of the other languages. We find that automated translation, particularly with GPT-4, holds substantial promise in enabling crosslingual analysis of cybercrime forums.
- We evaluate the effectiveness of repurposing English-language classification models for multilingual cybercrime data via translation, and compare their performance against models operating directly on the native language (§VI). We observe that classification performed on translated data is as effective as classification on the original language text.
- We explore few-shot and zero-shot learning strategies over multilingual underground forum data (§VII). Fully automated classification works but performs worse than human-annotated approaches.

We conclude by discussing implications and limitations (§IX).

II. Background

Understanding and analyzing CTI content from cybercrime communities is of great interest to analysts, law enforcement, and academics to understand new forms of criminal behavior [9]. However, this requires dedicated tools and methods. We describe key aspects of these methods.

Creating Ground Truth. Classification tasks require ground truth for training (supervised methods) or validating (unsupervised methods), typically in the form of labeled posts for underground forum analysis. The labeling method depends on the classification task. In some cases, posts can be automatically labeled with knowledge extracted from third parties (e.g., OSINT information for *artifacts* contained in a post [6]), which have typically gone through an annotation process supervised by humans. As a result, most researchers rely on manual labeling by human experts (e.g., to determine the crime type of a post [5]). Label confidence is usually obtained through the inter-agreement of multiple annotators. However, labeling is a resource-intensive task that requires experts with profound domain knowledge and native-level fluidity in the posts' language. These requirements make it difficult and costly to find experts.

Text Content Features. Classifying forum content often relies on features derived from written content, such as text-based features. General patterns can be captured using character counts and n-grams, while semantics and intent require embeddings from advanced NLP methods. These provide a more thorough representation of the textual content, putting them in the right context to grasp the nuances of the conversations. However, the success of these advanced NLP methods depends on pretrained models with adequate training data. Domain-specific models (e.g., a model trained on forum data) can have an advantage when applied to specific tasks. However, processing underground forum text is more difficult than well-formed formal text, as it contains jargon and slang, technical language and code, as well as noise and spelling mistakes. These limitations add difficulty for NLP models to evaluate the text, and require models trained for this task [28].

Classification Tasks. Various methods have been developed to classify content on underground forums. In particular, existing frameworks support targeted analysis by automatically inferring cybercrime-related characteristics from posts, such as predicting crime types or the presence of Indicators of Compromise (IoCs). However, most works focus on English-language classification, overlooking significant cybercrime activity in other linguistic domains that constitute major cybercrime hubs, such as Russian and Spanish. Besides existing automated evaluation methods, in many disciplines, such as criminology or psychology, researchers rely on manual qualitative analysis of forum posts. Due to the vast amounts of data, this can be slow, work-intensive, and limited to the languages the researchers understand. This highlights the need for reliable translation methods for cross-disciplinary research [29].

Research Gap. Current methods for analyzing underground forums generally do not handle multilingual content effectively. There are several areas where existing frameworks struggle. First, obtaining ground truth requires human experts with adequate domain knowledge and native-level fluidity. Such resources are scarce in less common languages, making their analysis less cost-effective compared to English. Second, as the extraction of post-content features is often language-specific, methods based on these features are not readily transferable to other languages. Third, while classification tasks still depend primarily on human annotation, the scarcity of domain experts presents a major obstacle, especially for researchers analyzing multilingual data.

While numerous domain-specific NLP models exist for English, comparable resources are often lacking for other languages, particularly within the nuanced context of cybercrime ecosystems. Processing content in less widely spoken languages remains challenging [28], and there is currently no established approach for handling multilingual underground forum data. Both the research and CTI communities commonly rely on MT to analyze such content; yet, it remains unclear whether MT preserves the domain-specific signals necessary for accurate classification. Recent advances in multilingual language models offer promising—but underexplored—alternatives for cross-lingual analysis. In particular, it is not yet known whether language-specific models can effectively extract informative features for classification, or how their performance compares to traditional classifiers applied to translated text.

III. Multilingual Text-based Classification

To understand multilingual cybercrime data, we set out to answer the following Research Questions (RQs):

RQ1: **Translating domain-specific multilingual data for humans.** What is the best approach to translate multilingual data while preserving content and meaning for cybercrime research?

RQ2: **Classifying multilingual data using translation and existing English-language models.** Can we repurpose prior work with English-language classification models for use with translated texts, or does the performance drop require language-specific models?

RQ3: **Multilingual few and zero-shot classification.** Can we side-step existing classification approaches to directly classify multilingual data using LLMs?

To answer these questions, we design an assessment pipeline depicted in Figure 1. Our methodology has three main steps, tailored to each RQ. First, we rely on known metrics and recruit native-speaker domain experts to determine the best-performing translator (§III-A). A well-performing translator not only facilitates the annotation of multilingual data in the presence of a language barrier, but also serves as a foundation for adapting prior work developed initially for another language, typically English. To assess whether translation is a reliable enabler for machine processing, the next step of our methodology (§III-B) compares the performance of a classification in the original language with that of its translation. This comparison allows us to assess whether translation-based adaptation maintains sufficient classification quality, or whether training language-specific models is necessary to avoid performance degradation. The results help to inform best practices for multilingual NLP in domains like cybercrime, where labeled data and resources are often concentrated in English. Finally, we investigate if a fully (zero-shot learning) or semi-fully (few-shot learning) language-agnostic analysis is possible (§III-C), assuming a language model pre-trained with sufficient contextual information around cybercrime.

A. Translating Multilingual Data

To study the suitability of MT for interpreting multilingual data, we assess the extent to which various MT systems preserve content and meaning relevant to cybercrime research from the perspective of an analyst. To achieve this, we perform a comparative analysis using a dataset comprising multiple underground forum posts in different languages. We translate posts from their original language to English using various translation systems, encompassing both proprietary and open technologies. To facilitate translation using LLMs, we design custom prompts tailored to each model inspired by the structure proposed in previous work [18], illustrated in Appendix A-A. We assess the quality of the translation from three axes. First, we perform an *atomic* assessment of the quality of MT, leveraging linguistic and semantic cues, to determine if they can be used as a proxy for assessing the quality of translation in underground forums, just as they are used in other contexts. Second, we resort to native speakers with certified fluency in English as a means to obtain reliable ground truth. Finally, we investigate whether the best translation can be evaluated automatically.

1) **Atomic assessment of linguistic and semantic cues:** We apply a combination of cross-lingual analysis techniques to capture both linguistic and semantic consistency. First, we perform a Name Entity Recognition (NER) and extract the Part of Speech (POS) tags of both the original text and its translation. We then compare these two sets using the cosine similarity and use this measure as a means to automatically determine which MT system works best [30]. Second, we use automated translation quality evaluation without referenced translation: BERTScore [31] and NMTScore [32]. BERTScore uses a neural network-based metric that measures semantic similarity using pretrained language models. We use this metric due to its performance in general-purpose benchmarks [33]. NMTScore [32] uses neural machine translation (NMT) models to estimate the likelihood that one sentence is a good translation of another.

2) **Human Assessment:** We recruit native speakers to evaluate the quality of the translations. Our recruits fulfill two key requirements: (1) they are domain experts, and (2) have proficient English language fluidity. Notably, due to the scarcity of human annotators meeting these two requirements, and the scale of languages evaluated (Arabic, German, Russian, Spanish, and Vietnamese), we assign one annotator per non-English language. However, we ensure that the assessment is done without attribution (i.e., we ensured assessors did not know which model produced each translation), and over a significant number of posts.

3) **Translation Evaluation with LLMs:** Lastly, we evaluate LLMs' effectiveness in assessing translation quality, building on prior work showing promising results [34], [35]. For this, we prompted the original text together with all the translations to different LLMs and tasked them to rank the quality of the translation. Our translation prompt emphasizes that translations require a deep understanding of the source language, including its slang and colloquial terms in areas like hacking, as illustrated in Appendix A-B.

B. Classifying Multilingual Content

This step of our methodology directly addresses whether existing English-language classification models can be effectively repurposed for multilingual cybercrime data through translation. Specifically, we evaluate the extent to which MT, followed by English-based NLP processing, can replicate the classification performance achieved by native-language NLP pipelines. To do this, we apply two established classification tasks originally developed for English-language data: i) identifying the type of cybercrime discussed in a post, and (ii) detecting posts containing malicious IoCs. We chose these two tasks because they represent recent and widely studied classification challenges that are highly relevant to the research community [9], [6]. We then compare their performance across multiple languages using two pipelines: one where native texts

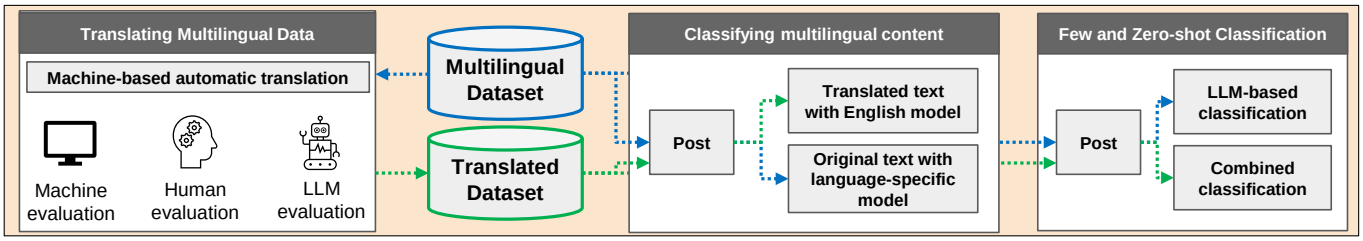

Fig. 1. An overview of the methodology.

are machine-translated into English and processed with the original English-language models, and another where native texts are processed directly using available language-specific tools. Next, we describe in detail the methods used to build the two classification tasks.

1) Crime Type Detection: We apply the classification pipeline of Atondo Siu et al. [5] to automatically detect the type of crime discussed in multilingual posts. This classification task enables a more targeted analysis of underground forum posts. This pipeline relies on keyword-based lexical features, represented by TF-IDF vectors, alongside XGBoost for prediction. As the original classification task is designed for English data only, we apply it to the *to-English* translated data. Furthermore, we adapt the language-specific NLP steps of the pipeline to support multiple languages, specifically regarding word tokenization and stopwords, and apply this modified pipeline to our dataset in its original language.

Importantly, we annotate posts in their original language to construct a reliable dataset suitable for multilingual classification tasks. Due to the need for language-specific expertise and quality assurance, our analysis is limited to two languages: Spanish and German, for which we had access to qualified annotators. Each post is labeled according to its associated crime type, following the taxonomy proposed by Atondo Siu et al. [5], from which both the labels and their descriptions are adapted with the modification that `Not criminal` has been adjusted to `None`—meaning, none of the other crime types detected. This is done to include posts discussing criminal activity not related to any of the given cybercrime categories. To aid interpretation and maintain consistency across annotators, each crime type is accompanied by an anonymized example. For completeness, the crime type labels and examples are shown in Appendix B.

Given that this dataset serves as the foundation for downstream classification tasks, where each post's label plays a critical role, ensuring annotation reliability is essential. To minimize labeling errors and resolve discrepancies, we employ multiple annotators and a structured conflict resolution strategy, and provide them with additional training. Initially, two annotators independently assign one or more crime-type labels to each post, allowing for multilabel annotations. The label is retained as a final annotation when there is consensus on at least one label. For posts with no overlapping labels between the two annotators, a third annotator independently re-annotates the post. In such cases, the final label set is determined by the intersection between the third annotator's labels and those of either of the initial annotators. We exclude posts where consensus is not reached.

Finally, we investigate how the labeling performance changes when performed on the translated data using the best translator evaluated in §III-A.

2) IoC Prediction: We extend the methodology in [6] by evaluating the impact of multilingual sentence-embedding models on classification performance. We depart from a task that performs targeted IoC detection by classifying forum posts as malicious or non-malicious based on the textual content of the post. This classification approach relies on *post-content features*, specifically sentence embeddings generated from English-translated content using the monolingual model *"all-mpnet-base-v2"* [36]. We apply the same pipeline in two settings: (1) using English-translated posts with the original monolingual embedding model, and (2) using the original-language posts with a multilingual sentence-embedding model. Here, ground truth labels are derived from OSINT sources as in [6]. A post is labeled as malicious if it contains at least one IoC that is flagged as malicious by OSINT, and as non-malicious if it contains only IoCs verified as non-malicious. Posts containing both non-malicious and unknown IoCs are excluded from the analysis to maintain label reliability.

C. Few and Zero-shot Classification

We investigate whether a classification task can be fully automated, to leverage it as a substitute for human annotators in generating ground truth, given the limited availability of domain experts for non-English languages. A notable advantage of LLMs is their ability to operate without requiring task-specific training data. Leveraging the generalization capabilities of LLMs, we employ both few-shot and zero-shot strategies for this task. While fine-tuning can help improve the performance of translation [18], it remains unclear whether fine-tuning also helps for labeling. To this end, we investigate whether the accuracy of labeling improves with an example per label (few-shot classification) or if zero-shot classification (without examples) is sufficient. To leverage the capabilities of the LLM, we apply prompt engineering techniques as outlined in [18] (see Appendix A-C). The prompt replicates the instructions and contextual information given to human annotators, ensuring consistency across labeling approaches.

We follow two alternative approaches. The first approach performs end-to-end classification, where the LLM is tasked to classify all posts in the evaluation set. The second approach performs a *2-step machine based classification*. In this setting, the LLM is used to annotate a subset of posts, which then serves as training data for a conventional classification pipeline

over the same evaluation set. We assess the quality of the two approaches by comparing their performance against human-annotated ground truth in the original language, thereby determining the extent to which LLMs can replicate human-level annotation performance.

IV. Dataset and Evaluation

Our multilingual dataset is derived from CrimeBB, the largest maintained collection of underground forums in multiple languages [2]. CrimeBB is available through data sharing agreements with the Cambridge Cybercrime Centre.[1] We use an existing dataset to avoid the time-consuming process of crawling new data. For our analysis, we include *all* available non-English language sources in CrimeBB. Specifically, we study the following five languages: Arabic, German, Russian, Spanish, and Vietnamese. We attempted to include data from AZSecure, a repository that claims to contain data from other languages, such as Chinese [37]. Unfortunately, this repository is no longer maintained and the data is not publicly available.

Translation. To evaluate the best translator, we create a dataset `<transet>`. We pick 400 posts each from all five available languages. Our sampling criteria aimed to capture both recent and a wide range of discussion topics, while maintaining an adequate content length for scaling the labeling. The post should also be understandable without reading the entire thread (i.e., without considering replies to previous comments). Hence, with a pick of 400 posts per language, the following rules apply: posts should be first in the thread, have a length of 20 to 300 words, and be from recent years (2021-2024); only German posts can be older, as the forum was closed in 2022. Additionally, we select at least 10 posts per subforum, where possible, and then randomly choose posts across all subforums until we have completed the 400 posts per language.

Crime-Type Labeling. For the crime-type labeling evaluation, we create a dataset `<crimeset>` consisting of 1,638 Spanish and 1,773 German posts. As described in §II, the labeling requirements made it difficult to find at least three experts for the remaining original languages to guarantee reliable labels, so we only include these two languages. Labeling is done according to §III-B1. For six German and three Spanish posts, there was no label intersection after three rounds of annotations. Therefore, those posts are excluded.

IoC Prediction. To evaluate this classification task, we create a dataset `<iocset>`. Following the pipeline of [6], we use IoC-Searcher [38] to extract *artifacts*. Specifically, we searched for 25k posts (5k per language) containing at least one *artifact* type (Domains, URLs, IPs, SHA1, SHA256, and MD5). We exclude those containing the forum domain (i.e., self-links). For the remainder, we scan the *artifacts* with VirusTotal [39] and use the scan results to characterize the posts as malicious or benign, discarding the posts whose *artifacts* are unknown to the antivirus industry. Hence, our dataset contains 11,160 posts. Figure 2 shows the breakdown by language and label.

[1]https://www.cambridgecybercrime.uk/data.html

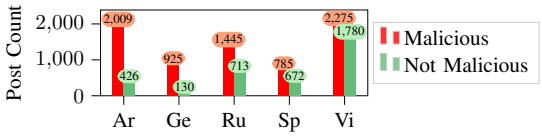

Fig. 2. Amount of posts with malicious (IoCs) and not-malicious *artifacts* per language: (Ar)abic, (Ge)rman, (Ru)ssian, (Sp)anish and (Vi)etnamese.

Evaluating our RQs. We utilize these datasets to address our research questions. In particular, (1) we use `<transet>` to ascertain whether translation meets human standards (RQ1), (2) we use `<crimeset>` and `<iocset>` to understand if translation is good enough for machines (RQ2), and (3) we use the same datasets as in the previous step to understand if we need humans at all (RQ3). The following sections describe our setup, results, and findings.

V. Translating Multilingual Data (RQ1)

We investigate the performance of four different translators when applied to `<transet>`:

- *GPT-4*, a state-of-the-art LLM developed by OpenAI [40], known for its strong capabilities across a wide range of NLP tasks, including translation.
- *Google Translate*, a commercial system using NMT models, continuously updated with large-scale data and user feedback [41].
- *DeepL*, another commercial translation system recognized for fluency and linguistic nuance, particularly in European languages [42].
- *MistralAI 7b_instruct_v0.2*, a free, open-weight LLM released in 2024, known for its competitive performance in several NLP benchmarks [43].

We explore MistralAI as the first three systems operate as paid services and are standard tools for translation. Despite MistralAI being an open-weighted model, the literature reports comparable translation quality to commercial systems [44]. The LLM prompts are inspired by previous works [18]. MistralAI, in contrast, requires a more elaborate and explicit prompt to achieve satisfactory results, likely due to differences in instruction-following capabilities.

We translate the `<transet>` dataset using all four translation models (excluding Vietnamese with DeepL, as it was not supported at the time of writing). Next, we present the evaluation results for the translators. First, we explore the potential of automated machine-based methods for reliable translator selection in the absence of human judgments and then assess a ranking provided by native human annotators. We also explore the capability of LLMs to rank the translations.

A. Atomic Assessment of Translation Quality

We calculate the cosine similarity between the number of NER and POS tags extracted from both the original texts and translations. To obtain named entities, we use the *xlm-roberta* model from Meta, finetuned for the NER task.[2] To extract morphological text features, we use the Stanza library from

[2]https://huggingface.co/FacebookAI/xlm-roberta-large-finetuned-conll03-english

the StanfordNLP group [45]. To use automated translation quality evaluation without referenced human translation, we use BERTScore—neural network-based metrics that evaluate semantic similarity using pretrained language models, which demonstrates best-performance in recent benchmarks [33]. Finally, we calculate the NMTScore with the SMALL-100 model, which supports 101 languages.[3] We utilize the results of the NER model to detect three types of entities: Person, Location, and Organization. A similarity of 100 means that the same model can extract the same entities in both the original and the translated text. We present our findings in Table I. GPT-4 outperforms other systems in translating German, Russian, and Spanish—not with an uplifting similarity. Google Translate achieves the best results for Arabic, while Mistral shows superiority for Vietnamese.

TABLE I
NAMED ENTITIES RECOGNITION RESULTS

model	Ar	Ge	Ru	Sp	Vi
DeepL	0.3846	0.2196	0.2587	0.0594	-
Mistral	0.3008	0.2365	0.2992	0.0640	**26.65**
Google Translate	**0.4256**	0.2769	0.2759	0.640	0.1243
GPT-4	0.3068	**0.2966**	**0.2994**	**0.0655**	0.1208

In the task of POS-tagging, Mistral performs best for Arabic and Russian, while GPT-4 performs best for German and Vietnamese, as shown in Table II. DeepL yields similar results to Mistral for Arabic, and the best results for Spanish. Google Translate performs as well as Mistral for Russian. POS-models detected 17 universal POS tags, while for our experiment we used only the five most informative tags, i.e., verbs, nouns, adjectives, adverbs and numbers.

TABLE II
PART OF SPEECH TAGGING RESULTS

model	Ar	Ge	Ru	Sp	Vi
DeepL	**0.2750**	0.2385	0.2385	**0.9225**	-
Mistral	**0.2750**	0.2385	**0.3640**	0.9201	0.2072
Google Translate	0.2699	0.2514	**0.3640**	0.9191	0.2020
GPT-4	0.1060	**0.2973**	0.2750	0.9195	**0.2073**

We furthermore evaluate translation by measuring semantic similarity between the original and translated text [46], as reported in Table III. We calculate BERTScore [31] with the use of multilingual transformer embeddings using *xlm-roberta-large*. Google Translate performs slightly better for Arabic, Russian, and Vietnamese, while DeepL shows slightly higher performance for German and Spanish. Finally, Table IX (in Appendix C) presents the NMTScore evaluation results, which align with the BERTScore findings. Google Translate shows superior performance for Arabic, Russian, and Vietnamese, whereas DeepL performs better for German and Spanish.

We observe consistent results when using metrics based on transformer embeddings, such as BERTScore and NMTScore: Google Translate demonstrates better performance for Arabic, Russian, and Vietnamese, while DeepL outperforms others

[3]https://huggingface.co/alirezamsh/small100

TABLE III
BERTSCORE (F1)

model	Ar	Ge	Ru	Sp	Vi
DeepL	0.9041	**0.5833**	0.9037	**0.5894**	-
Mistral	0.8756	0.4677	0.9151	0.4377	0.8982
Google Translate	**0.9052**	0.5697	**0.9262**	0.5755	**0.9169**
GPT-4	0.9037	0.5610	0.9224	0.5617	0.9088

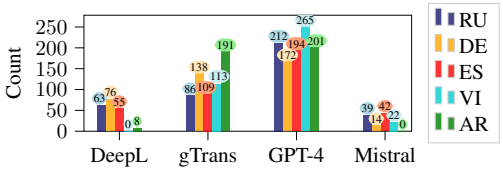

Fig. 3. Translation ranking according to human experts.

for German and Spanish. These findings partially support recent studies indicating that DeepL tends to perform better for European languages [47], [48]. In contrast, our evaluation based on the similarity of automatically extracted linguistic features from source and translated texts shows less consistent patterns. In the NER analysis, GPT-4 translations show better performance for Indo-European languages (German, Russian, and Spanish). The POS analysis revealed that GPT-4, Mistral, and DeepL each performed best in two different languages.

Overall, the lack of a consistent top performer across evaluations suggests that automated, general-purpose metrics alone are *inadequate* for selecting the best translation system for cybercriminal forum texts. The next section presents an expert evaluation to provide deeper insight.

B. Human Evaluation of Translation Quality

Figure 3 presents the frequency with which each translator was ranked as the top performer for each target language, based on post-level evaluations conducted by human experts who identified the best-performing model per translation. The primary evaluation criteria are the preservation of the original meaning and intent, correct handling of proper nouns (e.g., product names), and accurate interpretation of abbreviations and keywords. Additionally, technical terminology and code syntax must remain unaltered.

We find GPT-4 demonstrates the strongest ability to preserve original meaning. It effectively translates keywords and abbreviations and is robust to grammatical and spelling errors, which can lead to improved readability in some cases. Google Translate exhibits similar behavior, although at a slightly lower level of performance. Occasionally, it truncates Arabic texts. DeepL produces translations that are close to the original text, employs an advanced vocabulary, and maintains structural mistakes such as missing or double commas. This occasionally hinders the overall clarity and quality of the translation. In contrast, MistralAI is the least reliable in manual evaluations. It occasionally truncates or adds content, processes text incompletely, introduces incorrect word substitutions, alters or removes syntax, and often fails to convey the intended meaning and nuance. Additionally, instead of translating, it interprets or explains the prompt. Despite these issues, MistralAI can sometimes produces accurate translations and grammatical

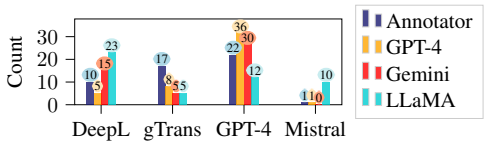

Fig. 4. Frequency of a translator being selected as top choice by LLMs.

corrections. While it introduces noise into quantitative translations, in a qualitative evaluation setting refined prompting can lead to reliable results. In general, all models demonstrate high translation quality, though *none* consistently handle slang appropriately. When the source text is poorly structured or contains numerous errors, all models experience difficulty.

C. AI-based Evaluation of Translations

To investigate whether LLMs can be used to rank the translation quality, we prompt GPT-4, LLaMA 3-8B-Instruct[4] and Gemini 1.5 Pro[5] to rank the translations for 50 sentences per language. The prompt is listed in Appendix A-B. Figure 4 depicts how often each model or human ranked a translator as top choice (rank 1). We note that LLaMA was unable to produce responses for three posts in Arabic. Human annotators ranked GPT-4 as the top translation 22 times. Compared to this baseline, LLMs showed skewed confidence. Gemini and GPT-4 over-prefer GPT-4 translations, selecting them 30 and 36 times respectively, while LLaMA showed lower preference, selecting it only 12 times. LLaMA chooses Mistral as the best model 10 times more than human annotators and DeepL about twice as much. All models overlook Google Translate as a second-best translator. These inconsistencies with human judgments suggest that LLMs are currently *not* reliable for assessing whether translations of underground forum posts accurately preserve the original meaning, intent, and context.

RQ1 Takeaway. Our evaluation indicates GPT-4 provides the most reliable translations across all five tested languages—Arabic, German, Russian, Spanish, and Vietnamese. Furthermore, it occasionally enhances the quality of posts by correcting minor spelling errors and accurately interpreting abbreviations and domain-specific keywords. However, none of the automated evaluation methods we examined—including cross-lingual analysis techniques and LLM-based assessments—consistently aligned with human judgments of translation quality.

VI. Classifying Multilingual Content (RQ2)

Given GPT-4's performance relative to humans, our next step is to assess whether semantic integrity is preserved when such translations are used in downstream multilingual classification tasks. Specifically, we examine whether high-quality translations are necessary for English-language models to perform effectively on translated content. Thus, we evaluate the reliability of machine translation and language-specific NLP tools within two established classification approaches, exploring whether translation can effectively repurpose English-based models for multilingual data.

[4]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[5]https://ai.google.dev/gemini-api/docs/models#gemini-1.5-pro

A. Crime-Type Labeling

This experiment compares the classification of `<crimeset>` in the original language (*olang*) to the classification of `<crimeset>` translated using GPT-4 (*trans-gpt*), each using its own tokenizer. We use a 60:20:20 split for training, testing, and validation. As shown in Table IV, F1-scores are very similar for both experiments, with Spanish performing slightly better in the *olang*, and German with *trans-gpt*. However, this could be due to the distribution of labels in the training and testing datasets, as 303 samples were used for testing Spanish posts and 337 samples for German posts. The size and distribution of the testing dataset resulted in the German dataset missing `ddos_booting` and contained 352 of 377 samples of `currency_exchange`. The Spanish dataset contained a wider distribution of labels: 157 `systems_access`, 107 `ddos_booting`, 21 `other`, 14 `bots_malware`, and 2 of `vpn_hosting` and `spam`.

TABLE IV
Comparing Original Language to GPT-4 Translations — Crime Type Prediction Using Ground Truth

Training Set	Precision	Recall	F1-Score	Accuracy
Spanish *olang*	0.86	0.86	**0.85**	0.86
Spanish *trans-gpt*	0.84	0.85	0.84	0.85
German *olang*	0.90	0.93	0.91	0.93
German *trans-gpt*	0.94	0.94	**0.92**	0.94

For this pipeline, we process *olang* Spanish and German individually as the NLP models and tokenizers only support one language at a time. Unifying the language through translation will eliminate this additional effort. This experiment suggests that the performance difference when processing data in *olang* or *trans-gpt* is negligible when classifying crime type. To further validate our finding, one author independently labeled 300 translated posts in each language. These labels were then compared to those assigned by native-language annotators, yielding an overall agreement of 93% (89.6% for Spanish and 96.3% for German). This high level of consistency, in line with the results reported in Table IV, reinforces the reliability of our results.

B. Predicting IoCs

For our comparison, we stick to the original pipeline described in [6], with a few modifications. Some *Metadata-based features* are not available for all forums in `<iocset>`. Given *Metadata-based features* account for only 6.8% of the overall feature importance [6], their exclusion has a negligible impact on the analysis. Instead, we focus on *Post-content features* and *Text-based features*, which collectively contribute the remaining 93.2%. For the comparison, the posts are translated with *trans-gpt*, while sentence-embeddings are generated with the model *"all-mpnet-base-v2."* To obtain the sentence-embeddings for our *olang* dataset, we use the

model *"multilingual-e5-large-instruct."* For the classification, we split the posts randomly into a training set (80%) and a test set (20%), using the same split for *olang* and *trans-gpt* for comparability. The results, shown in Table V, suggest the analysis over the *olang* performs slightly better than over the translated text, with a difference in F1-score of less than 0.01. For *trans-gpt*, it finds more true positives at the cost of flagging more false positives. Overall, the performance difference between *olang* and for *trans-gpt* is also negligible.

TABLE V
PERFORMANCE OF IOC-PREDICTION

	Accuracy	Precision	Recall	F1-Score
olang data	0.86	0.9	**0.89**	0.89
trans-gpt data	0.84	0.87	0.89	0.88

RQ2 Takeaway. We find the differences in processing data in *olang* (original language) or *trans-gpt* (GPT-4 English translation) are negligible. For crime type prediction, the Spanish dataset showed an increased performance of ≈0.01 F1-score when processed with *trans-gpt*, whereas for German the *olang* performed better by ≈0.01. For IoC prediction processing, *olang* showed an increased reliability by an F1-score of ≈0.01. It follows that when using GPT-4 as a translator, comparable results can be achieved as when processing the data in the original language. Naturally, this might slightly change if a different translator is used. However, our findings suggest we *can* rely on translated data to extract CTI when using GPT-4 for translation. This reduces the dependency on language-specific tools and enables research in low-resource languages where high-quality NLP tools are not available. It enables more efficient processing through a simplified architecture. Moreover, standardization of pipelines ensures consistency in model behavior and, therefore, results.

VII. FEW AND ZERO-SHOT CLASSIFICATION (RQ3)

The crime type classification task still relies on human annotation for ground truth. We investigate a further step of classification facilitation, where we build a classification pipeline not dependant on human annotation.

A. Classification with LLM

We select GPT-4 to perform our labeling task on `<crimeset>`. The prompt, derived from the GPT-4 translation prompt of §III-A, is shown in Appendix A-C. It contains the same instructions and information provided to the human annotators for labeling, i.e., annotator guidelines, subforum title, thread title, and post content, as well as an example for each crime type [5]. We assess both the few-shot classification by passing GPT-4 the crime type labels with examples, as well as the zero-shot classification (without examples). The labels, their description, and the examples are listed in Appendix B. As the translation *trans-gpt* performed best according to our experiments in §V,we will use those to translate subforum title, thread title, and post content for each post, and then compare the labeling performance of GPT-4 for each post. We label: 1) *olang* few-shot; 2) *olang* zero-shot; 3) *trans-gpt* few-shot; 4) *trans-gpt* zero-shot. The results are shown in Table VI.

TABLE VI
LABELING PERFORMANCE OF GPT-4 ON DIFFERENT SETTINGS: DATASET TYPE FEW-SHOT (FS) OR ZERO-SHOT (ZS)

Setting	Precision	Recall	F1-Score	Accuracy
olang FS	0.82	0.81	**0.81**	0.96
olang ZS	0.81	0.81	0.81	0.96
trans-gpt FS	0.81	0.80	0.81	0.96
trans-gpt ZS	0.81	0.79	0.80	0.96
olang FS Spanish	0.74	0.73	0.74	0.95
olang FS German	0.89	0.88	0.89	0.98

We observe that all classification settings generally perform well with correctly detected labels of ≈80%. The performance decreases by <1% when performed on translated text. This indicates that the labeling on the original text is preferred, as little performance is lost when working with the translation. The *Spam* category is hardest to identify in every setting (3.12%-4.69% agreement). A closer investigation reveals that all missed posts are in the Spanish dataset, where a large thread discusses *"propagation through Facebook"*. GPT-4 classified them as `none` or `bots_malware`, showing the model has difficulties with the spamming intention. Furthermore, it over-identifies `identity_theft` in all four settings (265.57%-300% more mentions; making 162-183 instead of 61 posts), mostly in the German dataset. A deeper inspection reveals that about ≈33% of those posts were also subject to labeling conflict among the human annotators, where one labeled the posts as `identity theft`, while the other labeled them as `none`. Finally, the third annotator sided with `none`, resulting to a disagreement with GPT-4. This illustrates how subtle the line can be when labeling such posts. The setup *olang* FS is the best performing one, therefore we also investigate the difference in the languages. The Spanish testing set achieves an F1-score of 0.74, while the German set achieves an F1-score of 0.89. A closer investigation reveals that 240 of the 438 misclassified posts were supposed to be `bots_malware`, but instead, 223 were classified as `none`.

With this setup, full GPT-4-based labeling is less reliable than conventional classification with a human-annotated training set (§VI-A), resulting in a decrease in Spanish F1-score of 0.11 and 0.03 in German. The performance gap also reflects the limitations of LLMs when applied to specialized domains like crime related classification tasks, where subtle semantic cues may be overlooked.

B. Scaling Automated Post Classification

To investigate a scalable classification approach, we label crime type through the *2-step machine based classification*. In this setting, we rely on the GPT-4 labeled dataset for training set, and use a conventional classification pipeline using XGBoost. For testing and validation, we rely on the labels obtained by human annotation, which serve as our ground truth. For a fair comparison with §VI-A, we use the

same dataset size (1.6k Spanish and 1.7k German posts) and split (60:20:20), and measure performance on both the *olang*-based and *trans-gpt*-based pipelines. We use language-specific German and Spanish tokenisation models for *olang*. Results are shown in Table VII.

Our results show that the proposed two-step classification approach works efficiently. F1-scores are identical in *olang* and *trans-gpt* for both languages, with 0.61 for Spanish and 0.88 for German. When compared to the human-annotated baseline, the German classification slightly decreases by an F1-score of 0.04. The performance is notably lower for Spanish, with a decrease in F1-score of 0.24. Note that German labels are majority 'None', and Spanish labels have a wider spread. Furthermore, the test set size is small, and better results could be expected with a larger dataset.

RQ3 Takeaway. These findings suggest while a fully automated machine-based approach is feasible, it falls short of the accuracy achieved through a human-based approach. In the *2-step machine based classification*, the lower quality labels of GPT-4 propagate errors into the downstream classifier, resulting in performance loss.

TABLE VII
COMPARING *olang* TO *trans-gpt* – CRIME TYPE PREDICTION WITH TRAINING LABELS FROM GPT-4 AND TESTING FROM HUMANS

Training Set	Precision	Recall	F1-Score	Accuracy
Spanish *olang*	0.63	0.67	0.61	0.67
Spanish *trans-gpt*	0.68	0.67	0.61	0.67
German *olang*	0.89	0.88	0.88	0.88
German *trans-gpt*	0.89	0.87	0.88	0.87

VIII. CASE STUDY

We presented a quantitative analysis of non-English cybercrime and reported a remarkable 93% agreement between annotations on original-language posts and their GPT-4-translated counterparts (cf. §VI-A). While this high level of consistency affirms the viability of translation-based pipelines, we present a qualitative examination of the remaining 7% of annotation disagreements between original and translated texts (RQ1 & RQ2) in §VIII-A, as well as of the differences between GPT-4 and expert annotators (RQ3) in §VIII-B.

A. Original vs. Translated Annotations

To better understand the sources of disagreement, we conduct a qualitative analysis focusing on the nature of annotation mismatches. Two main categories emerged from this analysis. The first includes posts related to *system access and hacking intent*, where discussions about tools or challenges may be interpreted as either legitimate or preparatory to cybercrime. The second involves posts on *bots, malware, or spam-related activity*, where content can be read as a neutral observation or as participation in illicit behavior. Next, we illustrate these disagreement types with representative examples and discuss the annotation challenges they present.

System Access and Hacking Intent. We observe several discrepancies regarding the intent of the hacking when examining the `systems_access` crime time. In *Example 1,* a person is asking in one post for help to set up a tool that is later used for hacking purposes. While the original language annotators labeled it as `none`, seeing the setup as non-criminal, the English annotator interpreted it as preparatory to a cybercrime and labeled it `systems_access`. In *Example 2,* an author shares a challenge about system penetration. Despite its educational framing, the description includes activities like gaining system access, privilege escalation, and decryption. This led the original language annotators to label it as `none`, but the English annotator classified it as `systems_access` based on potential implications.

Malware. In posts involving *bots, malware, or spam-related activity*, we see cases where the content can be interpreted either as a neutral observation or as active engagement in cybercriminal behavior, depending largely on how intent is understood. In *Example 3,* a user notes the presence of a bot active on the forum. The original language annotators categorized it as `bots_malware`, based on topic focus. However, the English annotator saw it as a neutral observation without direct engagement and labeled it as `none`. In another case, *Example 4,* an author asks if someone has a tool to spread and send large volumes of MSN emails. The English annotator labeled it as `bots_malware`, possibly due to the spreading mechanism, while the original annotators saw the spam-related goal and labeled it as `Spam`.

Summary. Disagreements between annotators stem *less* from issues of translation and *more* from the inherent ambiguity of online discourse in cybercrime forums—particularly in how intent is inferred from context. Posts that reference system access or hacking tools (Examples 1 and 2) often straddle the line between legitimate technical discussion and illicit intent. Similarly, posts involving malware (Examples 3 and 4) may appear either as neutral observations or as active participation in cybercriminal behavior, depending on how context is interpreted. These divergences highlight a key challenge in cybercrime annotation: the classification often hinges on inferred intent rather than explicit content. What one annotator sees as preparation for unauthorized access, another might interpret as harmless experimentation. Likewise, references to bots or spam tools may be read as descriptive or promotional depending on tone, phrasing, and assumed audience. *While transliteration ensures that the explicit content of posts is retained, annotators must still navigate the more nuanced challenge of interpreting intent—an inherently subjective process that extends beyond surface-level text.* Ultimately, this underscores the need for annotation frameworks that account for *intent, technical specificity, and platform norms*, and that support multi-label or probabilistic classifications where uncertainty is high.

B. Disagreement Between Annotators and GPT-4

To better understand the limitations of automated annotation, we conduct a qualitative analysis of cases where GPT-

4's labels diverge from those of human annotators. These disagreements often arise in ambiguous contexts that require nuanced interpretation, such as ambiguous illegalities, contexts, or labeling categories. By examining these examples, we highlight the challenges of context sensitivity that contribute to misclassification, offering insight into the boundaries of GPT-4's reliability in cybercrime annotation.

Ambiguous Illegality. We notice discrepancies in posts that describe ambiguous illegalities in the context of identity-related fraud. *Example 1:* In the German forum, users discussed how to receive parcels anonymously at packing stations, particularly by bypassing ID verification. Human annotators debated whether this behavior constituted cybercrime, ultimately labeling it as `none`, while GPT-4 classified it as `identity_theft`. *Example 2:* One post involves an author seeking a forum to arrange fake marriages, while another asks for a fake Bachelor's certificate. Though these activities suggest fraudulent intent, the human annotators labeled them as `none` due to their limited cybercrime relevance. GPT-4, however, classified both as `identity_theft`, likely overextending the crime category based on surface-level cues.

Context Misinterpretation. We see posts where context misinterpretation leads to mislabeling of the attack type. For instance *Example 3,* a post noting a service is back online and speculating a DDoS attack as the cause. GPT-4 labeled the post as `ddos_booting`, but since the post is speculative and lacks direct participation, the human annotators correctly classified it as `none`. We also see in a thread discussing malware, a user responds that a proposed setup would not function (*Example 4*). Without access to the surrounding conversation, GPT-4 labeled the response as `none`. However, human annotators, aware of the broader context, labeled it as `ddos_booting`, recognizing the technical discussion as part of a cybercrime.

Category Ambiguity. We also see posts involving spam or automation misclassified due to category ambiguity. For instance, several posts describe techniques for mass messaging on Facebook. Human annotators consistently labeled them as `Spam`, recognizing the intent to distribute unsolicited messages. GPT-4, however, misclassified them as `none` or `bots_malware`, potentially conflating spam with malware or overlooking the commercial spamming intent altogether.

Summary. The observed disagreements reveal systematic challenges in GPT-4's annotation behavior. The model often misinterprets the intent of legally or ethically ambiguous posts—especially when fraudulent behavior is discussed without a clear link to cybercrime. This issue is evident in cases of *Ambiguous Illegality*, where GPT-4 overextends labels like `identity_theft` to activities such as fake marriages or anonymous parcel pickups, which may involve deception but fall outside standard cybercrime taxonomies. Additionally, GPT-4 struggles with *Context Misinterpretation*, where accurate classification depends on surrounding discourse or prior messages. Without access to broader conversational threads, GPT-4 may mislabel posts that reference malware, DDoS attacks, or hacking setups, failing to detect whether the speaker is actively engaging in or merely commenting on a topic. A third source of disagreement lies in *Category Ambiguity*, where GPT-4 conflates overlapping or adjacent crime types—such as mistaking spam for bot activity or labeling neutral observations as active exploitation. These misclassifications underscore, again, a limitation in fine-grained categorization, especially when surface cues are weak or misleading.

Case Study Takeaways. Our first case study, comparing annotators in the original language and translation, suggests that disagreements do *not* stem from translation errors but from the *inherent ambiguity of certain posts*, where classification depends on context and interpretation. The second one, on disagreements between GPT-4 and experts, reveals that GPT-4 often misinterprets text, misses subtle cues, struggles with ambiguous posts, like those discussing fraud without explicit cybercrime intent. It also fails in understanding the need for more context.

IX. Discussions

By exploring strategies to process multilingual underground forums, our work derives critical findings with important implications for the research and CTI communities.

A. Findings

In this section, we first present the findings derived from evaluating each of our three research questions, and then discuss the broader implications that follow from our work.

RQ1: *What is the best approach to translate multilingual data for cybercrime research?* We find GPT-4 consistently produces the most reliable translations across five diverse languages, outperforming other MT systems according to human judgment (§V). For domain-specific data such as underground forum content, GPT-4 demonstrates strong potential as a translation solution, *enabling research on multilingual cybercrime data without the need for native speakers in each target language.* However, given the cost associated with using GPT-4 at scale, for qualitative analysis, open-source alternatives like Mistral can serve as viable substitutes, provided their outputs are subject to careful prompt engineering and human review. Research in LLMs remains a rapidly evolving landscape. As models continue to mature, improvements in translation quality, efficiency, and domain adaptability are likely to occur. In evaluating future translators, our findings reinforce that no automated approach currently matches the nuance and discernment offered by human judgment. Therefore, *human validation remains essential for assessing translation quality* in complex and specialized domains such as cybercrime analysis.

RQ2: *Can we repurpose prior work or do we require training language-specific models?* We find classification performance on GPT-4 translated data is comparable to processing data in the original language (§VI). This means existing English-language models and pipelines *can* be repurposed for translated data, *eliminating the need for language-specific adaptations in the context of our evaluation.* This significant finding enables scalable, standardized processing of multilingual

cybercrime data, especially in low-resource languages where (domain-) specific NLP tools are lacking.

RQ3: *Can we side-step approaches utilizing translation to instead directly classify multilingual data using LLMs?* The few- and zero-shot classification evaluation in §VII indicates that both a fully GPT-4-based classification approach and the *2-step machine based classification* method are effective for both translated and original-language data. While the end-to-end GPT-4 labeling achieves a competitive F1-score of 0.81 relative to the human-annotated baseline, its scalability is limited by the associated computational and financial costs. Our case study shows that GPT-4 misclassifies posts due to a lack of context as well as misinterpretation of subtle clues and ambiguous text. Moreover, labeling inaccuracies introduced by GPT-4 are further compounded when used in conjunction with conventional classification models, additionally degrading performance. Overall, the traditional pipeline, based on human-labeled data, consistently outperforms both the *2-step machine based classification* method and the fully automated GPT-4-based labeling approach. *This underscores that, for high-quality evaluations—particularly in domains such as cybercrime—expert annotations on the original language remain essential to reduce classification errors.*

Broader implications. We consider our three RQs collectively to address a broader question: *Given the difficulties of finding human annotators for certain languages, what alternative approaches can be employed to obtain annotations of comparable quality?* A central consideration in analyzing underground forum data is the tradeoff between annotation quality and resource investment, particularly the cost of domain experts with native-level fluency. While human annotation remains unmatched in quality, as shown in both our translation evaluation (§V) and few- and zero-shot classification experiments (§VII), relying exclusively on expert annotators is resource-intensive. Our results suggest that machine-based evaluation holds promise as a preliminary step, while a more cost-effective approach involves machine-based evaluation for pre-annotation or preliminary analysis, with humans refining the results. This hybrid strategy offers a scalable compromise, significantly reducing human workload while preserving the overall quality of results.

Our findings further show that classification performance is consistent whether the forum data is evaluated in its original language or in translation. This suggests that reliable MT enables effective annotation and classification without significant loss in performance. Importantly, this means domain experts no longer need to be native speakers of the target language, as accurate translation into English allows for high-quality human labeling. Our case study reinforces the findings derived in our quantitative analysis. This substantially lowers the barrier to involving qualified annotators and increases the feasibility of scaling up analysis across multiple languages. Moreover, we find key English-based classification models can be applied directly to translated data with minimal adaptation. This expands the applicability of existing tools and methods, making it easier to analyze underground forums in languages beyond English. Ultimately, our results support the feasibility of international and multilingual analysis efforts, which are critical in responding to cybercrime that routinely crosses linguistic and geographic boundaries [49].

Our work has two key takeaways. First, classification performed on translated underground forum data is as effective as classification on the original language text, indicating MT does not significantly degrade task performance. Second, for reliable, high-quality labeling and translation ranking evaluations, human expertise remains essential, as current machine annotations still fall short of human-level quality.

B. Limitations

Our findings are in the context of several limitations.

Post Coverage. Our study is based on a partial annotation of the CrimeBB dataset. Specifically, we randomly selected and annotated 2,000 posts for translation and labeled 3,411 posts, each reviewed by at least two annotators. While this represents a significant effort, the dataset for crime type labeling remains relatively limited in size compared to the full range and distribution of crime categories present in the corpus [5]. To mitigate potential bias, we employ stratified sampling and maintain consistent train, validation, and test splits across all experiments.

Language Coverage. We use the CrimeBB dataset, as justified in §IV. By including all available languages from CrimeBB, this evaluation encompasses a diverse range of languages spoken across Asia, North Africa, Europe, and South America. However, our findings might not generalize to languages not covered in this analysis, e.g., Chinese or Portuguese. Due to limitations in data availability, we are unable to obtain sufficient data from repositories containing these languages. Future research should aim to address this gap by collecting relevant data and applying the proposed methodology to a broader linguistic spectrum.

Reproducibility and Engineering. Our study relies on proprietary LLMs, which may limit the reproducibility of our work. However, we mitigate this by specifying the model version. Additionally, LLMs can generate distinct outputs in each run. To avoid this, and keep the output consistent, we set the temperature of GPT-4 to zero. Giving instructions to LLMs is often trial-and-error-based, task-specific, and model-dependent. Our study relies on carefully designed prompts to produce reliable translations, translation rankings, and labels from GPT-4. These prompts are derived from existing work targeting similar tasks, and outputs were manually tested. Improving the prompts could lead to better performance, in the same way that traditional classifiers can be parametrized to make models more efficient. However, our goal is not to develop industry-grade systems, but rather to advance the understanding of multilingual data analysis in the context of

cybercrime. The dataset is not public, but accessible through a data sharing agreement (cf. §IV).

X. Related Work

Machine Translation and Quality Evaluation. LLMs excel in zero- and few-shot translation tasks, outperforming traditional NMT systems in fluency and coherence, particularly for high-resource languages [50], [51]. Nonetheless, LLM translations often struggle with consistency in named entities, domain-specific jargon, and formal tone [52]. Automated metrics designed for sentence-level MT may not fully reflect the capabilities or limitations of LLMs [53]. Valeros et al. [18] provide initial insight into LLM-based translation of cybercriminal text. However, the study falls short in covering diverse, multilingual cybercrime datasets. It is limited to 100 sentences from one language, and the fine-tuned LLM approach does not scale as it requires ground truth in each input language. MT quality assessment remains important, as applications expand beyond general-purpose domains. A review of MT systems and quality assessment methods highlighted the dominance of Neural-based MT systems, and the need for more nuanced evaluation frameworks [54]. Traditional metrics such as BLEU often fail to capture semantic fidelity and contextual accuracy, particularly in domain-specific texts. Recent studies highlighted neural evaluation metrics like BERTScore [33], which better correlate with human judgments by leveraging transformer-based embeddings [55], [56]. Challenges remain in evaluating translations involving morphologically rich or low-resource languages. Kocmi et al. investigated the potential of LLMs for translation quality evaluation [34], as well as to detect translation quality errors with GPT-4 [35]. While their work is promising, their approaches have not been tested on cybercrime data.

Classification of Cybercriminal Text. Few works have addressed the problem of understanding multilingual posts in underground forums. Ebrahimi et al. propose the use of Adversarial Deep Representation Learning together, by applying transfer learning from the source language to English, detecting cyber threats in Russian, French [27], and Italian [8]. Different from our work, the authors proposed their own ML pipeline, instead of evaluating the capabilities of modern LLMs. Also, these works were designed to enhance machine classification, whereas in our work, we also aim to enable accessibility of information for humans through translation. Many existing works on underground forum data are only valid for English, limiting their applicability to other languages. For example, Caines et al. classify posts by author intent, post type, and addressee using statistical techniques combined with heuristics based on English keywords [14]. Pastrana et al. used a similar approach to identify key actors [15], and Atondo Siu et al. relied on an English-based tokenizer to classify posts by crime type [5]. Other works use more advanced NLP methods, supporting the English language only. Zhou et al. identify hate speech in hacking and extremist forum posts, using an English-specific sentence encoder [25], while Mischinger et al. developed an early IoC detection system for underground forum posts, relying on an English-based sentence transformer model that was applied to the English and *to-English* translated dataset [6]. DarkBERT is a language model pretrained on Dark Web data limited to the English language [28]. Works evaluating underground forums in multiple languages use less language-specific NLP methods like TF-IDF of character n-grams [7], [26], [57], thereby neglecting deeper textual understanding like meaning or syntax. Also, each language required distinct effort to process it (e.g. by investigation of language-specific keywords) [26]. For ground truth labeling, Bhalerao et al. [7] encountered difficulties in finding as many annotators for non-English text as for English.

XI. Conclusion

In this paper, we conducted an assessment of methods for processing multilingual data from underground forums. Our study, spanning five languages, constitutes the largest effort in the literature to date to understand non-English cybercrime forums. We compared four machine translation models, with human annotators identifying GPT-4 as the most reliable. Automated evaluation methods failed to determine the best model, highlighting the continued need for human judgment in MT assessment. Our study also shows that English-language classification pipelines can be effectively repurposed for translated multilingual cybercrime data. By leveraging high-quality machine translation, domain experts can annotate and classify content in non-English languages without requiring native fluency. This finding significantly lowers the barrier for scalable, standardized analysis of global underground forums and highlights the viability of translation-driven NLP approaches in multilingual, resource-constrained settings. Finally, while machine-generated annotations achieve reasonable performance, our findings show that reliance on automated labeling introduces notable errors that propagate through the classification pipeline, reducing overall effectiveness. This demonstrates that, despite advances in LLMs, machine-based annotation remains inferior to human labeling in high-stakes domains such as cybercrime. We conclude that human expertise remains indispensable for both translation evaluation and dataset labeling, and underscore the need for continued research into automating cybercrime analysis.

XII. Ethics

Ethics approval was granted from the department's ethics committee for this work. We used data collected from publicly available forums, and could not gain informed consent from all members as this would be considered to be spamming. As we only analyze posts as a collective whole, rather than identifying individual users, this falls outside of the requirement of informed consent. We also avoid publishing details that could identify individuals, including usernames and original post contents.

ACKNOWLEDGMENTS

We are grateful to Andrew Caines for organizing the Spanish and German annotations (supported by the Economic and Social Research Council (ESRC) (grant number ES/T008466/1)), as well as Anh V. Vu and Medhi Benatallah for the Vietnamese and Arabic annotations. Medhi Benatallah was supported by the King's College Summer Research Programme. This work was supported by the project PID2022-143304OB-I00 funded by MICIU/AEI/10.13039/-501100011033/ and by the ERDF, EU. Guillermo Suarez-Tangil is a 2020 RyC fellow RYC2020-029401-I, funded by MCIU/AEI/10.13039/501100011033 and the ESF Investing in your future. The same grant has funded Mariella Mischinger's work. Jack Hughes and Alice Hutchings are supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 949127). Sergio Pastrana was supported by grant PID2023-150310OB-I00 (MORE4AIO) of the Spanish AEI.

REFERENCES

[1] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "An analysis of underground forums," in *Proceedings of the 2011 ACM SIGCOMM Internet Measurement Conference*, 2011, pp. 71–80.

[2] S. Pastrana, D. R. Thomas, A. Hutchings, and R. Clayton, "CrimeBB: Enabling cybercrime research on underground forums at scale," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1845–1854.

[3] A. Bermudez-Villalva and G. Stringhini, "The shady economy: Understanding the difference in trading activity from underground forums in different layers of the web," in *Proceedings of the APWG Symposium on Electronic Crime Research (eCrime)*, 2021, pp. 1–10.

[4] S. Pastrana, A. Hutchings, D. Thomas, and J. Tapiador, "Measuring eWhoring," in *Proceedings of the Internet Measurement Conference*, 2019, p. 463–477. [Online]. Available: https://doi.org/10.1145/3355369.3355597

[5] G. Atondo Siu, B. Collier, and A. Hutchings, "Follow the money: The relationship between currency exchange and illicit behaviour in an underground forum," in *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2021, pp. 191–201.

[6] M. Mischinger, S. Pastrana, G. Suarez-Tangil *et al.*, "IoC Stalker: Early detection of Indicators of Compromise," in *Proceedings of the Annual Computer Security Applications Conference*, 2024.

[7] R. Bhalerao, M. Aliapoulios, I. Shumailov, S. Afroz, and D. McCoy, "Mapping the underground: Supervised discovery of cybercrime supply chains," in *Proceedings of the IEEE APWG Symposium on Electronic Crime Research (eCrime)*, 2019, pp. 1–16.

[8] M. Ebrahimi, S. Samtani, Y. Chai, and H. Chen, "Detecting cyber threats in non-English hacker forums: an adversarial cross-lingual knowledge transfer approach," in *Proceedings of the IEEE Security and Privacy Workshops (SPW)*, 2020, pp. 20–26.

[9] J. Hughes, S. Pastrana, A. Hutchings, S. Afroz, S. Samtani, W. Li, and E. Santana Marin, "The art of cybercrime community research," *ACM Computing Surveys*, vol. 56, no. 6, pp. 1–26, 2024.

[10] J. Ramírez Sánchez, A. Campo-Archbold, A. Zapata Rozo, D. Díaz-López, J. Pastor-Galindo, F. Gómez Mármol, and J. Aponte Díaz, "Uncovering cybercrimes in social media through natural language processing," *Complexity*, vol. 2021, pp. 1–15, 2021.

[11] M. Arazzi, D. R. Arikkat, S. Nicolazzo, A. Nocera, M. Conti *et al.*, "NLP-based techniques for cyber threat intelligence," *arXiv preprint arXiv:2311.08807*, 2023.

[12] J. Torregrosa, G. Bello-Orgaz, E. Martínez-Cámara, J. D. Ser, and D. Camacho, "A survey on extremism analysis using natural language processing: definitions, literature review, trends and challenges," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 8, pp. 9869–9905, 2023.

[13] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. R. B. Carvalho, and E. Stamatatos, "Authorship attribution for social media forensics," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 5–33, 2017.

[14] A. Caines, S. Pastrana, A. Hutchings, and P. J. Buttery, "Automatically identifying the function and intent of posts in underground forums," *Crime Science*, vol. 7, no. 1, pp. 1–14, 2018.

[15] S. Pastrana, A. Hutchings, A. Caines, and P. Buttery, "Characterizing Eve: Analysing cybercrime actors in a large underground forum," in *Proceedings of the 21st International Symposium on Research in Attacks, Intrusions, and Defenses (RAID)*, 2018, pp. 207–227.

[16] J. Lusthaus, M. Bruce, and N. Phair, "Mapping the geography of cybercrime: A review of indices of digital offending by country," in *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2020, pp. 448–453.

[17] M. Edwards, G. Suarez-Tangil, C. Peersman, G. Stringhini, A. Rashid, and M. Whitty, "The geography of online dating fraud," in *Workshop on technology and consumer protection*. IEEE-TCSP, 2018.

[18] V. Valeros, A. Širokova, C. Catania, and S. Garcia, "Towards better understanding of cybercrime: The role of fine-tuned LLMs in translation," in *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2024, pp. 91–99.

[19] D. Seyler, W. Liu, Y. Zhang, X. Wang, and C. Zhai, "Darkjargon. net: A platform for understanding underground conversation with latent meaning," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 2526–2530.

[20] K. Yuan, H. Lu, X. Liao, and X. Wang, "Reading thieves' cant: automatically identifying and understanding dark jargons from cybercrime marketplaces," in *Proceedings of the 27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1027–1041.

[21] Y. Li, J. Cheng, C. Huang, Z. Chen, and W. Niu, "Nedetector: Automatically extracting cybersecurity neologisms from hacker forums," *Journal of Information Security and Applications*, vol. 58, p. 102784, 2021.

[22] E. Vanmassenhove, D. Shterionov, and A. Way, "Lost in translation: Loss and decay of linguistic richness in machine translation," in *Proceedings of Machine Translation Summit XVII: Research Track*, 2019, pp. 222–232.

[23] A. Mukherjee and M. Shrivastava, "Lost in translation? found in evaluation: A comprehensive survey on sentence-level translation evaluation," *ACM Computing Surveys*, 2025.

[24] V. Ghafouri, J. Such, and G. Suarez-Tangil, "I love pineapple on pizza!= i hate pineapple on pizza: Stance-aware sentence transformers for opinion mining," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 21 046–21 058.

[25] L. Zhou, A. Caines, I. Pete, and A. Hutchings, "Automated hate speech detection and span extraction in underground hacking and extremist forums," *Natural Language Engineering*, vol. 29, no. 5, pp. 1247–1274, 2023.

[26] R. S. Portnoff, S. Afroz, G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, D. McCoy, K. Levchenko, and V. Paxson, "Tools for automated analysis of cybercriminal markets," in *Proceedings of the 26th International World Wide Web Conference*, 2017, pp. 657–666.

[27] M. Ebrahimi, Y. Chai, S. Samtani, and H. Chen, "Cross-lingual cybersecurity analytics in the international dark web with adversarial deep representation learning," *Mis Quarterly*, vol. 46, no. 2, 2022.

[28] Y. Jin, E. Jang, J. Cui, J. W. Chung, Y. Lee, and S. Shin, "Darkbert: A language model for the dark side of the internet," in *61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*. Association for Computational Linguistics (ACL), 2023, pp. 7515–7533.

[29] J. Hughes, Y. T. Chua, and A. Hutchings, "Too much data? opportunities and challenges of large datasets and cybercrime," in *Researching Cybercrimes: Methodologies, Ethics, and Critical Approaches*, A. Lavorgna and T. J. Holt, Eds. Springer, 2021, pp. 191–212.

[30] L. Han, "Machine translation evaluation resources and methods: A survey," *arXiv preprint arXiv:1605.04515*, 2016.

[31] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[32] J. Vamvas and R. Sennrich, "NMTScore: A multilingual analysis of translation-based text similarity measures," in *Findings of the Association for Computational Linguistics: EMNLP*, 2022, pp. 198–213.

[33] M. Hanna and O. Bojar, "A fine-grained analysis of BERTScore," in *Proceedings of the Sixth Conference on Machine Translation*, 2021, pp. 507–517.

[34] T. Kocmi and C. Federmann, "Large language models are state-of-the-art evaluators of translation quality," *arXiv preprint arXiv:2302.14520*, 2023.

[35] ——, "GEMBA-MQM: Detecting translation quality error spans with GPT-4," *arXiv preprint arXiv:2310.13988*, 2023.

[36] Hugging Face, "all-mpnet-base-v2," https://huggingface.co/sentence-transformers/all-mpnet-base-v2, 07 2022.

[37] S. Samtani, K. Chinn, C. Larson, and H. Chen, "Azsecure hacker assets portal: Cyber threat intelligence and malware analysis," in *2016 IEEE conference on intelligence and security informatics (ISI)*. Ieee, 2016, pp. 19–24.

[38] J. Caballero, G. Gomez, S. Matic, G. Sánchez, S. Sebastián, and A. Villacañas, "The rise of GoodFATR: A novel accuracy comparison methodology for indicator extraction tools," *Future Generation Computer Systems*, vol. 144, pp. 74–89, 2023.

[39] Virustotal, "Virustotal," https://www.virustotal.com/gui/home/upload, [Online] Last accessed: April, 30 2025.

[40] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[41] J. R. Jolley and L. Maimone, "Thirty years of machine translation in language teaching and learning: A review of the literature," *L2 Journal: An Electronic Refereed Journal for Foreign and Second Language Educators*, vol. 14, no. 1, 2022.

[42] Y. A. Telaumbanua, A. Marpaung, C. P. D. Gulo, D. K. W. Waruwu, E. Zalukhu, and N. P. Zai, "Analysis of two translation applications: Why is DeepL translate more accurate than Google Translate?" *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 4, no. 1, pp. 82–86, 2024.

[43] D. S. Chaplot, "Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed," *arXiv preprint arXiv:2310.06825*, 2023.

[44] A. K. Wassie, M. Molaei, and Y. Moslem, "Domain-specific translation with open-source large language models: Resource-oriented analysis," *arXiv preprint arXiv:2412.05862*, 2024.

[45] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

[46] H. Saadany and C. Orašan, "BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text," in *Proceedings of the Translation and Interpreting Technology Online Conference*, 2021, pp. 48–56.

[47] C. M. Hidalgo-Ternero, "Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation." Universitat d'Alacant, 2020.

[48] L. Matviienko, L. Khomenko, I. Denysovets, K. Horodenska, T. Nikolashyna, and I. Pavlova, "Comparative analysis of online translators in the machine translation system," *Revista Romaneasca pentru Educatie Multidimensionala*, vol. 16, no. 3, pp. 101–118, 2024.

[49] K. Huang, D. W. E. B. C. GrierD, T. J. Holt, C. Kruegel, D. McCoy, S. Savage, and G. Vigna, "Framing dependencies introduced by underground commoditization," in *Workshop on the Economics of Information Security*, 2015.

[50] W. Jiao, W. Wang, J.-t. Huang, X. Wang, S. Shi, and Z. Tu, "Is ChatGPT a good translator? Yes with GPT-4 as the engine," *arXiv preprint arXiv:2301.08745*, 2023.

[51] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li, "Multilingual machine translation with large language models: Empirical results and analysis," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 2765–2781.

[52] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla, "How good are gpt models at machine translation? a comprehensive evaluation," *arXiv preprint arXiv:2302.09210*, 2023.

[53] D. Elshin, N. Karpachev, B. Gruzdev, I. Golovanov, G. Ivanov, A. Antonov, N. Skachkov, E. Latypova, V. Layner, E. Enikeeva *et al.*, "From general LLM to translation: How we dramatically improve translation quality using human evaluation data for LLM finetuning," in *Proceedings of the Ninth Conference on Machine Translation*, 2024, pp. 247–252.

[54] I. Rivera-Trigueros, "Machine translation systems and quality assessment: a systematic review," *Language Resources and Evaluation*, vol. 56, no. 2, pp. 593–619, 2022.

[55] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7881–7892.

[56] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2685–2702.

[57] J. Burroughs, M. Tereszkowski-Kaminski, and G. Suarez-Tangil, "Visualizing cyber-threats in underground forums," in *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2023, pp. 244–258.

APPENDIX A
PROMPTS

We list the different prompts used in this work.

A. Prompts for Translation

We use two similar prompts for GPT-4 and Mistral in §III-A. The prompt is inspired by [18].

Prompt GPT-4 Figure 5 shows the prompt used for GPT-4.

```
'''You are a translator bot specializing in translating
source_language to EN.
You have deep knowledge of source_language, including slang
related to hacking, internet, military, finance, and vulgar
or colloquial terms.
Do not translate names of websites, URLs, services, media,
or companies. Keep names consistent in English.
Translate dates, links, and informal slang while preserving
the original tone.
Do not modify syntax.
Do not explain the translation, just provide it.
Ensure accuracy and context-appropriate translations as per
these guidelines.
Here is the text: text'''
```

Fig. 5. Prompt used for GPT-4 translation.

Prompt MistralAI Figure 6 shows the prompt used for MistralAI.

```
'''You are a Language Translator Bot specialized in
translating from source_language to English.
You have a deep understanding of source_language.You deeply
understand source_language slang related to hacking,
internet, network attacks, military terms, military
equipment, financial terms related to money, loans, and
lending, and vulgar, offensive and colloquial words.
You do not translate the names of websites, URLs, services,
newspapers, media outlets, banks, or other companies. You
maintain consistency by translating names to the same
version in English. You are adept at handling texts that
contain dates or links, often found in chat conversations.
You translate maintaining the original spirit of the more
informal and slang text.
Do not modify syntax.
You do not explain the translation.
You only write the translation.
Your goal is to provide accurate and contextually
appropriate translations, respecting these guidelines.
Here is the text:
text
'''
```

Fig. 6. Prompt used for MistralAI translation.

B. Prompt for Translator ranking

Figure 7 is the prompt used for the translation ranking in Section III-A3. This prompt is inspired by the prompt used for translation in Section III-A and by [34] and [35].

C. Prompt for Few- and Zero-shot Classification

Figure 8 shows the prompt used for labeling the posts in §III-C. The prompt is based on the prompt for translation used in §III-A and adapted for the labeling task. Depending on whether the labeling is done as few-shot (with examples) or zero-shot(without crime type examples) classification, the dictionary {crime_type} contains labels and explanations or additionally an anonymized example after each explanation.

```
''' (System) You are an annotator for the quality of machine
translation from source_language to target_language. You
have deep knowledge of source_language and target_language,
including slang related to hacking, internet, military,
finance, and vulgar or colloquial terms.
There is one original text and four different translations,
each surrounded with triple backticks. Your task is to rank
the four translations from 1(best) to 4(worst).
A good translation preserves the meaning, grammar and
undertone and understands keywords and slang, while keeping
the syntax close to the original. A bad translation is not
preserving the meaning, adding or removing text or changing
syntax.
Do not explain the ranks, just provide them.
Write your answer in the form rank1: translator_name \n
rank2: translator_name \n rank3: translator_name \n rank4:
translator_name

original text: ```row["original"]```\n
translation_d: ```row["translation_deepl"]```\n
translation_g: ```row["translation_gpt"]```\n
translation_m: ```row["translation_mistral"]```\n
translation_t: ```row["translation_gTrans"]``` '''
```

Fig. 7. Prompt used for ranking the translations.

APPENDIX B
CRIME TYPE LABELS

Table VIII lists the crime type labels, their description, and the anonymized example given to the authors in §III-B1 and to GPT-4 in §VII.

APPENDIX C
TRANSLATION EVALUATION

Table IX presents the NMTScore of our translation evaluation in §V-A.

TABLE VIII
CRIME TYPE LABELS WITH THEIR DESCRIPTION AND EXAMPLE

Label	Description	Anonymized example
systems_access	Access to systems (excluding use of malware) and SQL injection attacks.	How to access a phone's text messages and calls without physical access to it.
bots_malware	Bots or malware and related services.	How to make my server file (of RAT) FUD????
eWhoring	eWhoring (simulation of fraudulent cybersexual encounters for financial gain).	I am new to eWhroing. Can someone please gimme some tips/advice? PM me for my Skype.
currency_ex	Exchanging digital currencies.	Looking for Amazon.de giftcards
ddos_booting	DDoS attacks, booting, stressing, and stress testing.	Would you be interested in investing in a SST service 100% money would be made back plus more.
identity_theft	Online identity theft, internet fraud, online scams or credit card fraud.	I want to buy a Ebay USA sms account verification service
spam	Sending spam, sharing email addresses or containing marketing services.	Earn passive money with clickbank
trading_creds	Trading accounts including gaming, social networks and Netflix accounts.	Selling sickest kik
vpn_hosting	VPN, hosting and proxy services.	I am looking for someone to host OMCPool.net in return for a share in the profits.
none	None of the above crime types detected.	

```
''' You are a labeling bot specializing in labeling the
crime type of forum posts in source_language.
You have deep knowledge of source_language, including slang
related to hacking, internet, military, finance, and vulgar
or colloquial terms.
Do not explain the labels, just provide them.
The crime type can be one of the following labels, further
explained in a dictionary that contains the labels as keys
and their explanation as values:
crime_type = {crime_type}
Multi-labels are allowed for all annotation categories. Only
pick the exact label names provided in the dictionary.
Write your answer together with a justification for your
choice in the form label1
or for multiple labels label1,label2,label3
Some posts discuss crime types but not the actual commission
of the crime being discussed. For example, one of the posts
on the bulletin board called 'Suggestions and Ideas' asks
for 'Death Removal of eWhoring'. The discussion focuses on
whether the topic of 'eWhoring' should or could be removed
from the forum. These types of posts are classified as 'not
criminal'.
Some posts discuss products or services, but do not indicate
they are selling or using them. For example, some posts
discuss software that can automatically increase forum
participants' social network channels' subscribers by the
hundreds. Where there is no indications the products were
being used or sold, these posts are classified as 'not
criminal'.
Posts about 'Botting' and 'Hosting' can be particularly
difficult to classify as they are not always related to
criminal activity. For example, the use of bots for enhanced
game playing is a common topic. While this may be against
a game's terms of service, no crime is being committed.
Therefore, analysing the post context is crucial for
categorising the post correctly. As such, these types of
posts are classified as 'not criminal'.

Ensure accuracy and context-appropriate labeling as per
these guidelines.
You have information about the post, the thread and the
sub-forum th post appears in.
This is the sub-forum title: {subforum_title}
This is the general thread headline: {thread_title}
This is the post content: {text}
'''
```

Fig. 8. Prompt used for labeling forum posts with crime types.

TABLE IX
NMTSCORE.

model	Ar	Ge	Ru	Sp	Vi
DeepL	16.91	**27.64**	26.53	**30.87**	-
Mistral	11.85	19.64	21.62	18.65	14.07
Google Translate	**17.63**	26.78	**26.60**	28.99	**21.27**
GPT-4	15.66	24.50	22.83	26.85	16.61