MIS Quarterly

RESEARCH NOTE

CROSS-LINGUAL CYBERSECURITY ANALYTICS IN THE INTERNATIONAL DARK WEB WITH ADVERSARIAL DEEP REPRESENTATION LEARNING[1]

Mohammadreza Ebrahimi
School of Information Systems and Management, University of South Florida,
Tampa, FL 33620 U.S.A. {ebrahimim@usf.edu}

Yidong Chai
School of Management, Hefei University of Technology,
Hefei, Anhua 230009 CHINA {chaiyd@hfut.edu.cn}

Sagar Samtani
Department of Operations and Decision Technologies, Indiana University,
Bloomington, IN 47405 U.S.A. {ssamtani@iu.edu}

Hsinchun Chen
Department of Management Information Systems, University of Arizona,
Tucson, AZ 85721 U.S.A. {hchen@eller.arizona.edu}

International dark web platforms operating within multiple geopolitical regions and languages host a myriad of hacker assets such as malware, hacking tools, hacking tutorials, and malicious source code. Cybersecurity analytics organizations employ machine learning models trained on human-labeled data to automatically detect these assets and bolster their situational awareness. However, the lack of human-labeled training data is prohibitive when analyzing foreign-language dark web content. In this research note, we adopt the computational design science paradigm to develop a novel IT artifact for cross-lingual hacker asset detection (CLHAD). CLHAD automatically leverages the knowledge learned from English content to detect hacker assets in non-English dark web platforms. CLHAD encompasses a novel Adversarial deep representation learning (ADREL) method, which generates multilingual text representations using generative adversarial networks (GANs). Drawing upon the state of the art in cross-lingual knowledge transfer, ADREL is a novel approach to automatically extract transferable text representations and facilitate the analysis of multilingual content. We evaluate CLHAD on Russian, French, and Italian dark web platforms and demonstrate its practical utility in hacker asset profiling, and conduct a proof-of-concept case study. Our analysis suggests that cybersecurity managers may benefit more from focusing on Russian to identify sophisticated hacking assets. In contrast, financial hacker assets are scattered among several dominant dark web languages. Managerial insights for security managers are discussed at operational and strategic levels.

Keywords: Cybersecurity analytics, dark web, automated hacker asset detection, cross-lingual knowledge transfer, adversarial learning, computational design science

[1] H. R. Rao was the accepting senior editor for this paper. Dmitry Zhdanov served as the associate editor.

DOI: 10.25300/MISQ/2022/16618

Introduction

Cybercrime is estimated to cost the global economy $6 trillion annually by 2021 (Morgan, 2017). A large portion of cybercrime stems from the dark web (Chen, 2012), a conglomerate of international and ever-evolving online platforms mainly characterized by hacker forums and dark net markets (DNMs) (Pastrana et al., 2018). The dark web is rife with malicious assets that hackers can leverage to launch attacks that compromise the cybersecurity of individuals and organizations. Hacker assets include malware, hacking tools (e.g., phishing/carding tools), hacking tutorials (e.g., procedures to monetize stolen credit cards), and malicious source code (Benjamin et al., 2019). Dark web content is a valuable cybersecurity resource as it provides reconnaissance on hacker assets. These assets reflect hackers' tools, techniques, and procedures (TTP) and provide a unique opportunity to understand adversaries' arsenals and capabilities.

The popularity of dark web platforms among cybercriminals has led to an increase in the number of illicit items from several thousand in 2013 to hundreds of thousands in 2018 (Dittus et al., 2018). Given the magnitude of the dark web, it is impractical for human analysts to manually sift through the content to identify hacker assets. However, automatically detecting hacker assets among thousands of other similar illegal items (e.g., pirated e-books, digital goods) is a nontrivial task. Keyword-based searching approaches are prone to inclusion and exclusion errors (Ebrahimi et al., 2020). Recognizing this issue, recent cybersecurity reports suggest utilizing automated machine learning (ML) techniques to monitor the dark web for hacker assets (Tolido et al., 2019). While ML approaches hold significant promise in automating hacker asset detection, their training procedures require human-labeled data, which is expensive and time consuming to obtain (Portnoff, 2018). This issue becomes more pronounced when performing hacker asset detection in foreign languages. The language barrier makes acquiring human-labeled training data more expensive for non-English dark web platforms (Ebrahimi et al., 2018). ML models' performance often suffers in low-resource environments that lack human-labeled data (Tian et al., 2018). Current cybersecurity analytics studies attempt to alleviate this issue by using machine translation (MT) services to translate non-English (low-resource) data to English where more labeled training data exists (high-resource) (Samtani et al., 2017). However, MT services are trained on general-purpose corpora from the web (e.g., Wikipedia articles), and are often not suited for translating domain-specific languages (Devlin et al., 2019; Johnson et al., 2017). Moreover, the hacker-specific language in the dark web is rife with jargon causing mistranslations that affect hacker asset detection performance (Yuan et al., 2018). Today, Russian, French, and Italian are among the most common languages in the dark web (Schäfer et al., 2019). Nation-specific dark web platforms differ in the type of hacker assets they host (Benjamin & Chen, 2016). Thus, analyzing non-English content helps security analysts and others better understand the global cybersecurity landscape (Schäfer et al., 2019). Accordingly, scholars have emphasized the critical need for multilingual dark web cybersecurity analytics research (Benjamin et al., 2019). One promising approach for responding to this need is to leverage knowledge from human-labeled English content to analyze low-resource non-English content, known as cross-lingual knowledge transfer (CLKT). Against this backdrop, we adopt the computational design science paradigm (Rai, 2017) to develop a novel CLKT framework, cross-lingual hacker asset detection (CLHAD), to automatically detect hacker assets in non-English dark web platforms without MT. At the core of CLHAD stands a novel adversarial deep representation learning (ADREL) method. Drawing upon state-of-the-art methodologies in generative adversarial networks (GANs), ADREL is a novel method that automatically extracts language-invariant representations from English contexts and transfers them to non-English contexts without requiring external resources (e.g., human- or machine-translated corpora) or extensive human-labeled training data. Rather than relying on lexicons to translate words while ignoring their context, ADREL constructs representations from textual descriptions of dark web hacker assets that embed salient features from the source and target language. Our study contributes to cybersecurity analytics by presenting a novel multilingual hacker asset detection framework (CLHAD with ADREL) and conducting multilingual hacker asset profiling on the international dark web. Based on the results of this profiling, security managers may benefit more from focusing on Russian platforms in identifying sophisticated hacking assets. In contrast, identifying financial hacker assets requires attending to several dominant languages in the dark web.

Research Background

Three areas of literature are examined. First, we explore IS cybersecurity literature to position our work within cybersecurity analytics, the overarching area for hacker asset detection. Second, we review the prevailing cybersecurity analytics methods for automated hacker asset detection in the dark web. Third, we review CLKT literature to guide the transfer of knowledge from English to non-English content as an effective approach to reduce human labeling costs. Based on the reviews, we identify research gaps and the research problem we study in this work.

IS Cybersecurity Literature and Cybersecurity Analytics

IS cybersecurity studies can be categorized by their paradigms into behavioral, economic, and computational design science. Recent studies in the behavioral paradigm focus on modeling user behavior in cybersecurity, including habituation to security warnings (Vance et al., 2018), user training for mitigating phishing attacks (Jensen et al., 2017), and operational risk management (S. Yang et al., 2017). Studies in the economic paradigm focus on modeling the economic impact of security phenomena (e.g., the impact of law enforcement and discussions in hacker forums on DDoS attacks; Hui et al., 2017; Yue et al., 2019) and the impact of diversifying software resources on security risk (Temizkan et al., 2017). The emergence of high-impact and publicly available dark web data (as opposed to proprietary data sources) has given birth to a stream in IS cybersecurity research that focuses on *cybersecurity analytics* (Chen et al., 2012). Studies in this stream often adopt the computational design science paradigm to develop novel IT artifacts that provide ML models for automated decision-making (Benjamin et al., 2019; Li et al., 2016; Yin et al., 2019). These studies construct models for various analytical cybersecurity tasks such as de-anonymizing cybercriminals (Yin et al., 2019), classifying hacker assets (Ebrahimi et al., 2020; Samtani et al., 2017), and identifying key hackers (Benjamin et al., 2016; Li et al., 2016). However, these studies either offer monolingual models (e.g., English only) or rely on MT. The rapid growth of the international dark web calls for novel cybersecurity analytics methods capable of digesting multilingual content to support global automated hacker asset detection.

Cybersecurity Analytics for Automated Hacker Asset Detection in the Dark Web

Most online platform analytics studies focus on well-structured English content culled from well-known social media platforms such as Twitter and Facebook (Shore et al., 2018). In contrast, dark web platforms contain a plethora of unstructured, non-English content and hacker jargon that challenge the applicability of the existing analytics approaches to hacker asset detection. Consistent with Benjamin et al. (2019), hacker assets can be identified as *hacking tools* that are software designed to circumvent security controls and illicitly manipulate technologies (e.g., remote access Trojan, carding tools), *malicious source code* (e.g., uncompiled executables or scripts), and *hacking tutorials* that provide instructions to hackers for executing specific tasks (e.g., stealing cryptocurrency). Non-hacker assets have limited cybersecurity relevance or value and include digital goods (e.g., illicit multimedia), copyrighted software, pirated e-books, counterfeits, drugs, forged documents, and others. Table 1 summarizes recent automated hacker asset detection research in the dark web in two key dimensions: the examined language and approach to analyze non-English content.

Prevailing hacker asset detection methods leverage ML algorithms such as SVM and K-means (Marin et al., 2016; Nunes et al., 2016; Portnoff, 2018), deep learning (Grisham et al., 2017; Queiroz et al., 2019), topic modeling (Deliu et al., 2017), and keyword search (Benjamin & Chen, 2015). Most studies support only English content due to the high cost of non-English human-labeled data, which could be attributed to language barriers. Also, data labeling remains an issue in studies that examine non-English content by constructing separate monolingual models for each language (Duong et al., 2016). To alleviate this issue, some studies employ MT to leverage the English labeled data for non-English content. However, MT errors can lower the performance of automated hacker asset detection for two major reasons. First, prevailing MT services are trained on general surface web documents such as books, news, and Wikipedia articles (Devlin et al., 2019) or European Parliament records (Johnson et al., 2017), for which human-translated versions exist. Thus, they are not designed for domain-specific, low-resource applications such as dark web cybersecurity analytics (Schäfer et al., 2019), where the hacker-specific language is laden with evolving jargon that causes numerous word order and semantic translation errors (Benjamin & Chen, 2015). Second, MT services use sequence learning models that require high-quality human-translated documents as ground truth (Cao & Xiong, 2018). Obtaining multilingual translation training corpora is far more expensive than human-labeled data for text classification (Duek & Markovitch, 2018). CLKT is a promising approach to address these challenges by transferring known hacker asset knowledge obtained from high-resource English platforms to non-English ones (Rasooli et al., 2018). To our knowledge, our proposed framework offers the first multilingual hacker asset framework that effectively operationalizes CLKT without relying on machine translation.

Cross-Lingual Knowledge Transfer (CLKT)

CLKT draws upon a branch of ML known as transfer learning. Transfer learning aims to leverage knowledge from a resource-rich source domain to solve a task in a target domain that lacks sufficient training data (Weiss et al., 2016). CLKT aims to improve learning a target task in a low-resource language by using the knowledge acquired from a high-resource language. Traditional CLKT approaches require a carefully engineered set of features to transfer. However, feature engineering is a manual, *ad hoc*, labor-intensive process that needs significant domain knowledge.

Table 1. Selected Recent Studies on Automated Hacker Asset Detection in the Dark Web

Year	Author(s)	Task	Method(s)	Platform	Language(s)	Approach
2020	Ebrahimi et al.	Identifying hacker assets	TSVM	7 DNMs	en	Monolingual
2019	Queiroz et al.	Identifying hacker assets	CNN	3 forums, 1 DNM	en	Monolingual
2018	Portnoff	Classifying malicious posts	SVM	8 forums	en, ru, de	Monolingual
2018	Deliu et al.	Identifying hacker assets	SVM, LDA	1 forum	en	Monolingual
2017	Deliu et al.	Identifying hacker assets	CNN	1 forum	en	Monolingual
2017	Samtani et al.	Classifying malware code	SVM, LDA	8 forums	en, ru	MT
2017	Grisham et al.	Detecting mobile malware	RNN	4 forums	en, ru, ar	MT
2016	Nunes et al.	Identifying hacker assets	SVM	10 DNMs	en	Monolingual
2016	Marin et al.	Grouping malicious products	K-means	17 DNMs	en	Monolingual
2015	Benjamin et al.	Identifying hacker assets and vulnerabilities	Keyword search	5 forums, 4 carding shops	en, ru	Monolingual

Note: CNN: Convolutional Neural Network; LDA: Latent Dirichlet Allocation; SVM: Support Vector Machine; TSVM: Transductive SVM; RNN: Recurrent Neural Network; en: English; ru: Russian; de: German; ar: Arabic.

Moreover, the extracted features are often context-specific and may not generalize.

In light of these issues, recent studies have adopted deep learning methods for CLKT. Rather than transferring manually-constructed feature sets, deep architectures automatically extract salient language representations. Deep CLKT has been used extensively for tasks such as foreign Twitter message classification (X. Yang et al., 2017), multilingual sentiment analysis (Dong & de Melo, 2018), and multilingual speech recognition (Ning et al., 2017). However, CLKT relies on external resources, including multilingual embeddings, parallel corpora, and MT to facilitate knowledge transfer. Attaining these external language resources introduces three practical challenges. First, obtaining high-quality multilingual embeddings for domain-specific text is expensive (Li et al., 2017). Second, constructing parallel corpora is prohibitive due to the high cost of word/sentence alignment across the source and target languages (Abdalla & Hirst, 2017). Third, many languages lack reliable MT for domain-specific text (Abdalla & Hirst, 2017). These limitations affect the viability of conventional CLKT approaches for applications lacking these resources (e.g., dark web). Our proposed framework is uniquely positioned in the cybersecurity analytics literature as an approach that eliminates the need for these external resources by devising a novel GAN-based CLKT approach that learns language-invariant representations from the multilingual text.

In sum, within IS cybersecurity analytics research, approaches that leverage disparate multilingual data sources are lacking. Given the proliferation of international cybercriminal platforms, IT artifacts that enable analyzing both English and foreign-language dark web content are critically needed. Past hacker asset detection studies that use separate monolingual models suffer from data labeling issues (Queiroz et al., 2019). Studies that rely on MT suffer from mistranslations of hacker-specific language (Samtani et al., 2017). While CLKT can be helpful, it often relies on external resources (e.g., parallel corpora) to identify language-invariant features. Such resources are often costly to acquire or unavailable for dark web platforms. These research gaps motivate developing automated cross-lingual cybersecurity analytics for hacker asset detection within international dark web platforms. From the managerial perspective, past studies have highlighted the importance of automated multilingual hacker asset detection in obtaining holistic views of the global cybersecurity landscape (Ebrahimi et al., 2020; Samtani et al., 2017) and facilitating managerial tasks such as identifying hiring needs for security analysts with specific language proficiencies in security firms (Spataro, 2021). As such, CLHAD offers security managers a prescriptive way to holistically provide insights on the dark web in an explainable way.

Methodological Foundation

To inform our design, we first describe the role of deep representation learning as an effective approach to automatically extract language-specific salient features from text. Then, we explain GAN as a promising approach to provide language-invariant representations for CLKT.

Deep Representation Learning from Text

The promise of deep learning in text analysis is attributable to their ability to learn text representations that embed meaningful semantics captured through nonlinear transformations in each layer of the architecture (Bengio et al., 2013). These text representations reduce the dimensionality of text and are useful for downstream tasks such as text classification. Among deep architectures, bidirectional long short-term memories (BiLSTMs) are specifically designed to extract such representations from text. They leverage word order in forward and backward directions in text and capture temporally-dependent patterns, which are often missed by alternative deep architectures (Goldberg, 2017). This makes them suitable for extracting monolingual representations from text (Jozefowicz et al., 2016). However, the language-specific text representations generated by BiLSTMs are not transferable to other languages. Ganin et al. (2016) show that transferable representations must be language-invariant (i.e., common to both the source and target languages). GAN offers a deep generative model that can extract transferable, language-invariant features without external resources or manual feature engineering.

Generative Adversarial Networks (GANs)

GANs employ two neural networks that engage in an adversarial learning (AL) strategy (Goodfellow et al., 2016). AL is a game-theoretic approach for training two competing learning components (i.e., generator and discriminator) simultaneously. The generator (*G*) is trained to create "synthetized" data that are hard to discern from the "real" training data. The discriminator (*D*) learns to distinguish the real data from the synthesized data generated by *G*. AL can be formalized as a minimax game with a value function *V*:

$$\min_{G}\max_{D} V\left(D,G\right)=\mathrm{E}_{x}\left[\log D\left(x\right)\right] + \mathrm{E}_{z}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right] \tag{1}$$

where $x{\sim}P_{data}(x)$ is an instance from the distribution of the real-word training data and $z{\sim}P_z(z)$ denotes an instance drawn from the prior distribution of the input noise (e.g., uniform or Gaussian), and $D(x)$ is the probability that x came from real-world data. Equation 1 denotes the total discriminator's reward in the game. *G* is trained to minimize *D*'s reward, while *D* is trained to maximize its reward by assigning the correct labels to real and synthesized data. $\mathrm{E}_x[\log D(x)]$ denotes D's reward if it correctly predicts the real-world data is genuine. Similarly, $\mathrm{E}_z[\log(1 - D(G(z)))]$ represents *D*'s reward if it correctly predicts the generated data is synthesized. In theory, the game terminates when neither *G* nor *D* can improve further (i.e., GAN reaches equilibrium). GAN's AL procedure has four steps. In Step 1, the generator synthesizes initial samples from noise with a predefined distribution. In Step 2, the discriminator is trained to discern between the real and synthesized data. Step 3 compares the discriminator's prediction with the ground truth via a loss function and updates the generator's weights to improve the quality of synthesized data. Step 4 repeats Steps 1–3 until the generator synthesizes data indistinguishable from the real data (i.e., GAN equilibrium). While GANs have promise in facilitating CLKT without external resources, identifying how the generator and discriminator can interact to produce a language-invariant representation requires additional study.

Proposed Research Design

We propose a novel design to address the identified research gaps. Figure 1 shows the four major components of our research design: (1) data collection and pre-processing, (2) cross-lingual hacker asset detection (CLHAD), (3) performance evaluation, and (4) explanation and detected hacker assets profiling. Each component is detailed in the following subsections.

Data Collection and Data Pre-Processing

We identified and collected four large-scale hacker forums (one English, one Russian, and two French) and ten DNMs (seven English, one Russian, one French, and one Italian). These platforms are highly ranked in DeepDotWeb and DarkWebNews, two well-known directories regularly accessed by the cybersecurity community. We designed a Tor-routed, obfuscated crawler using breadth-first search (BFS) to extract product descriptions and post content and to parse them into a relational database. Our collection includes 862,715 items from 2016 to 2019, with 761,993 hacker forum posts and 100,722 DNM products. In total, 242,247 of the samples are English. The rest are Russian, French, and Italian. Consistent with past literature, our pre-processing steps include tokenizing the text, converting the characters to lower-case, and unifying the encodings to UTF-8. While the entire collection was used to learn text representations, a subset of the collection was used to construct the gold-standard set to train and evaluate the model (Li et al., 2016).

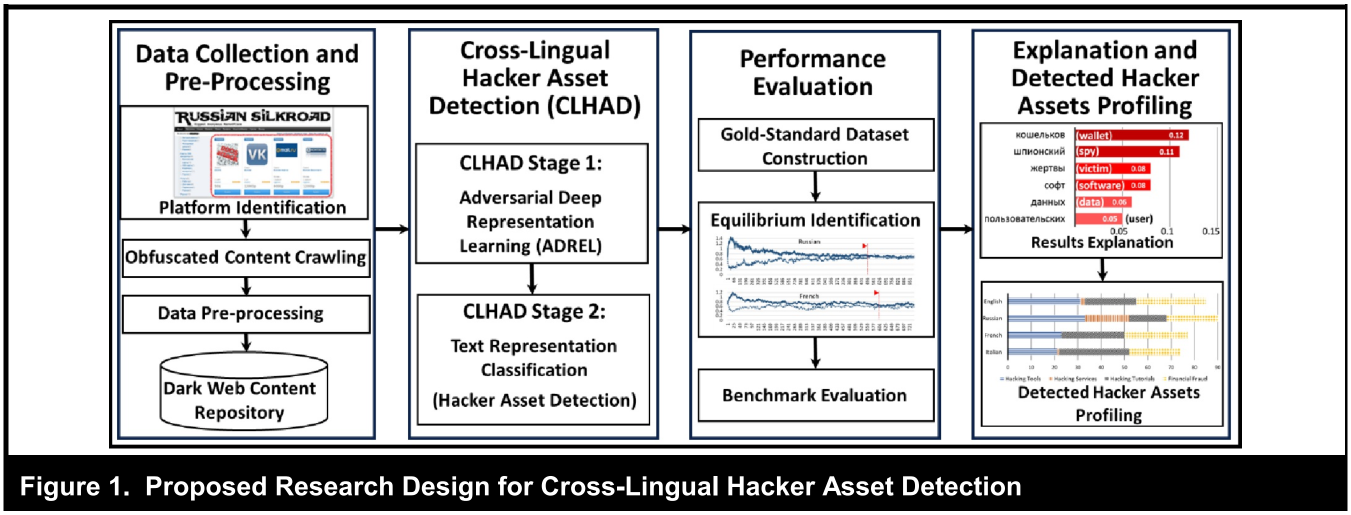

Data Collection and Pre-Processing
RUSSIAN SILKROAD
VK
Platform Identification
Obfuscated Content Crawling
Data Pre-processing
Dark Web Content Repository
Cross-Lingual Hacker Asset Detection (CLHAD)
CLHAD Stage 1: Adversarial Deep Representation Learning (ADREL)
CLHAD Stage 2: Text Representation Classification (Hacker Asset Detection)
Performance Evaluation
Gold-Standard Dataset Construction
Equilibrium Identification
Russian
French
Benchmark Evaluation
Explanation and Detected Hacker Assets Profiling
кошельков (wallet) 0.12
шпионский (spy) 0.11
жертвы (victim) 0.08
софт (software) 0.08
данных (data) 0.06
пользовательских 0.05 (user)
0.05
0.1
0.15
Results Explanation
English
Russian
French
Italian
0 10 20 30 40 50 60 70 80 90
Hacking Tools
Hacking Services
Hacking Tutorials
Financial Fraud
Detected Hacker Assets Profiling

Figure 1. Proposed Research Design for Cross-Lingual Hacker Asset Detection

Cross-Lingual Hacker Asset Detection (CLHAD)

Automatically extracting language-invariant text representations that are transferable from a high-resource language to a low-resource target language can aid hacker asset detection in non-English platforms. To this end, CLHAD comprises two stages: (1) a novel adversarial deep representation learning (ADREL) method to learn language-invariant representations from the textual content in English and non-English dark web platforms and (2) a binary classifier to categorize the learned representations as hacker assets. We detail each stage below.

CLHAD Stage 1: Adversarial Deep Representation Learning (ADREL)

The two-player game in AL facilitates constructing representations that contain essential features of two languages. Thus, we design ADREL to extract language-invariant text representations from source (English) and target (non-English) content. ADREL's training consists of two phases. First, language-specific representations are obtained from the source and target platforms via BiLSTMs. Second, language-invariant features are built from BiLSTM representations with a novel GAN design. In Phase 1, input word vectors from different languages are processed with separate BiLSTMs to generate language-specific representations. The English and non-English contexts are denoted by *en* and *NE*, respectively. Phase 2 learns language-invariant representations with an AL training approach. We present each phase in Figure 2.

The design intuition of our proposed AL strategy is that each language-specific representation (*en* or *NE*) needs modification to resemble the *opposite* language's representation for knowledge transfer. In this adversarial setup, the non-English generator learns to produce representations that carry the salient features from the English context for hacker asset detection. The generators are trained to mimic each other's representations, while the discriminator is trained to recognize the generated representation's language. ADREL's AL setup is formalized as the minimax game in Equation 2:

$$\min_{G^{en},G^{NE}} \max_{D} V\left(D,G^{en},G^{NE}\right) = \mathrm{E}_{R^{NE}}\left[\log\left(D\left(G^{NE}\left(R^{NE}\right)\right)\right)\right] + \mathrm{E}_{R^{en}}\left[\log\left(1-D\left(G^{en}\left(R^{en}\right)\right)\right)\right] \quad (2)$$

where R^{en} and R^{NE} are the initial language-specific text representations. G^{en} and G^{NE} are generators for English and non-English representations, respectively. Like conventional GAN, the equation denotes the total discriminator's reward in the game. Both G^{en} and G^{NE} are trained to minimize D's reward. D is trained to maximize its reward by correctly recognizing the language of synthesized data. $\mathrm{E}_{R^{NE}}\left[\log\left(D\left(G^{NE}\left(R^{NE}\right)\right)\right)\right]$ denotes D's reward if it correctly labels a synthesized representation coming from a non-English language. Finally, $\mathrm{E}_{R^{en}}\left[\log\left(1-D\left(G^{en}\left(R^{en}\right)\right)\right)\right]$ is D's reward if it correctly labels a synthesized representation coming from English. Both G^{en} and G^{NE} play against D until they learn to generate features that are common to both languages.

Conceptually, ADREL employs a "Y-shaped" architecture with two generators and one discriminator to set up the game-theoretic approach in Equation 2. Figure 3 compares ADREL's architecture and formulation with conventional GAN. As shown in Figure 3, ADREL modifies GAN to enhance the language-specific text representations (R^{en} and R^{NE})

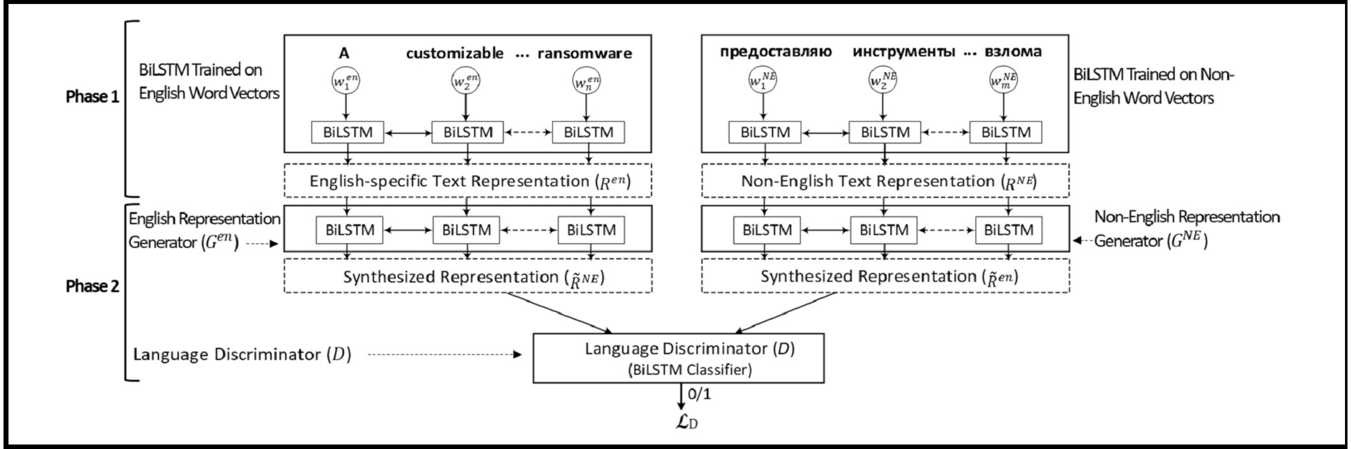

Phase 1
BiLSTM Trained on English Word Vectors
A
customizable
... ransomware
w_1^{en}
w_2^{en}
w_n^{en}
BiLSTM
BiLSTM
BiLSTM
English-specific Text Representation (R^{en})
предоставляю
инструменты
... взлома
w_1^{NE}
w_2^{NE}
w_m^{NE}
BiLSTM
BiLSTM
BiLSTM
BiLSTM Trained on Non-English Word Vectors
Non-English Text Representation (R^{NE})
Phase 2
English Representation Generator (G^{en})
BiLSTM
BiLSTM
BiLSTM
Synthesized Representation ($\tilde{R}^{NE}$)
Non-English Representation Generator (G^{NE})
BiLSTM
BiLSTM
BiLSTM
Synthesized Representation ($\tilde{R}^{en}$)
Language Discriminator (D)
Language Discriminator (D)
(BiLSTM Classifier)
0/1
$\mathcal{L}_D$

Figure 2. Adversarial Deep Representation Learning (ADREL) Architecture

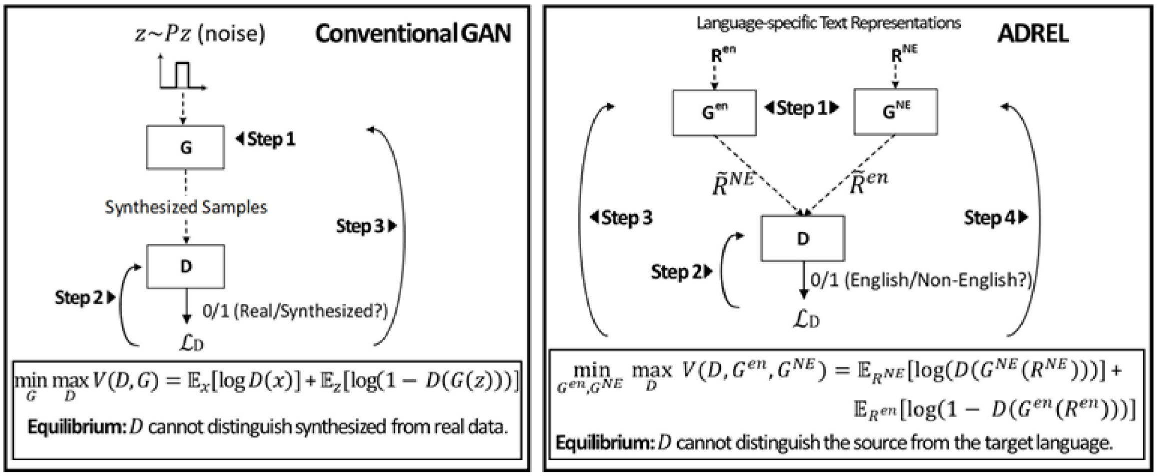

$z \sim Pz$ (noise)
Conventional GAN
G
Step 1
Synthesized Samples
Step 3
D
Step 2
0/1 (Real/Synthesized?)
$\mathcal{L}_D$
$\min_G \max_D V(D,G) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))]$
Equilibrium: D cannot distinguish synthesized from real data.
Language-specific Text Representations
ADREL
R^{en}
R^{NE}
G^{en}
Step 1
G^{NE}
$\tilde{R}^{NE}$
$\tilde{R}^{en}$
Step 3
Step 4
D
Step 2
0/1 (English/Non-English?)
$\mathcal{L}_D$
$\min_{G^{en},G^{NE}} \max_D V(D, G^{en}, G^{NE}) = \mathbb{E}_{R^{NE}}[\log(D(G^{NE}(R^{NE})))] + \mathbb{E}_{R^{en}}[\log(1 - D(G^{en}(R^{en})))]$
Equilibrium: D cannot distinguish the source from the target language.

Figure 3. Extending the Architecture and Formulation of the Standard GAN (left) to ADREL (right) for Learning Language-Invariant Representations

generated by the BiLSTMs from Phase 1. Generators aim to produce language-invariant features with a new AL formulation and a new equilibrium criterion where the discriminator D cannot distinguish the source language from the target language. To solve the minimax problem in Equation 2, ADREL employs a 5-step iterative process. Step 1 generates non-English and English synthesized data $\tilde{R}^{NE}$ and $\tilde{R}^{en}$ by applying generators G^{en} and G^{NE} on R^{en} and R^{NE}. Step 2 trains the discriminator D to maximize Equation 2 by recognizing the language of synthesized representations. Steps 3 and 4 train generator G^{en} and G^{NE}, respectively, to minimize Equation 2 by improving the quality of synthesized data given the feedback from D. Steps 1–4 repeat until neither G^{en} nor G^{NE} improves significantly (equilibrium). ADREL's novelty over prevailing CLKT and GAN methodologies is three-fold. First, ADREL utilizes two generators, each of which is encouraged by the same discriminator to create representations similar to that of the other generator. This setup creates language-invariant representations that can support downstream tasks with limited and costly labeled training data. Second, ADREL eliminates the need for external resources commonly seen in CLKT studies, including multilingual word embeddings (Tian et al., 2018), machine-translated corpora (Zhou et al., 2016), or parallel corpora (Xu & Yang, 2017). To the best of our knowledge, ADREL is the first method to extract language-invariant representations without requiring external language resources. Third, ADREL does not define a prior distribution on the data, and thus can be suitable for text applications since the text distribution is unknown. As a result, ADREL can support emerging applications for which there is limited prior knowledge. For reproducibility, our model specification and implementation is available at https://github.com/ebra-8/ADREL.

CLHAD Stage 2: Text Representation Binary Classification (Hacker Asset Detection)

The second stage in CLHAD aims to assign a label $\hat{y} \in \{0,1\}$ non-English text representation from stage 1, where 1 denotes hacker assets, and 0 indicates non-hacker assets. CLHAD outputs the assignment probability as a confidence score that can be useful for analysts to confirm or override the model's suggestions. After training ADREL, the resultant G^{NE} can be applied to the non-English data to extract the language-invariant features. A binary classifier can then be used to classify the text representations. Due to its strong performance in text classification, we used a BiLSTM binary classifier with a logistic loss function (Goldberg, 2017).

Performance Evaluation

To evaluate the performance of our proposed approach, we first established a gold-standard labeled dataset. This dataset served as the basis for two sets of evaluations. The first examined the quality of AL training via equilibrium identification. The second used the trained CLHAD model and compared its performance against prevailing ML and CLKT benchmark methods.

Gold-Standard Dataset Construction

To create a gold-standard testbed, we formed a panel of seven annotators. Five were native speakers (two Russian, two French, and one Italian), and the other two were cybersecurity experts. Each language was manually annotated by a native speaker and a cybersecurity expert. Initial briefing sessions were conducted for each language and platform. To assess the annotation's validity, we obtained an initial agreement rate between the native speakers and cybersecurity experts and identified the conflicting annotations. Additional meetings were conducted to discuss the conflicting annotations between the annotators. More than 99% of instances were unanimously annotated at this stage. Non-informative translations that could not be agreed upon were omitted, resulting in a gold-standard testbed with 5,976 total documents from four languages across 14 dark web platforms (3,271 English, 2,271 Russian, 713 French, and 435 Italian documents). This labeled dataset was used for training and evaluating ADREL and is available on our GitHub repository.

Equilibrium Identification

A common challenge in AL is identifying the appropriate number of training iterations needed to train GAN (Arjovsky et al., 2017), which in theory signifies reaching equilibrium. In practice, this is critical to ensure that the model is not under-trained so that high-quality representations are generated for downstream tasks. Monitoring generator and discriminator losses (i.e., errors with respect to the ground truth) during training can help identify the appropriate number of iterations (Arjovsky et al., 2017). The training iteration at which these losses start to stabilize can signify the effective training of GAN (Goodfellow et al., 2016). Figure 4 shows the losses at each training iteration for Russian (top), French (middle), and Italian (bottom). The stabilization points are marked with dashed vertical lines.

While the discriminator and generator losses exhibited an increasing trend in at the early stages of training, their losses reduced as the number of training iterations increased until they almost stabilized after about 500, 600, and 400 iterations for Russian, French, and Italian, respectively. These results guided the number of ADREL's training iterations in benchmark evaluations.

Benchmark Evaluation

We systematically evaluated CLHAD against state-of-the-art benchmark methods for hacker asset detection in Russian, French, and Italian as target, low-resource languages. A lexicon-based baseline and three sets of benchmark ML methods were identified from extant literature: (1) monolingual models, (2) MT-based, and (3) CLKT alternatives. The baseline used only non-English labeled training data and was based on searching hacker asset indicators from a crafted lexicon within a given text. The lexicon compilation process is detailed in our code repository. Monolingual models were trained only on non-English data without any translation to English or transfer of knowledge between languages. Two families of monolingual models were examined: (1) traditional ML models, including SVM, random forest, and naïve Bayes, and (2) monolingual deep learning models. Three deep learning models commonly used in past text classification literature were selected for evaluation: BiLSTM (Goldberg, 2017), bidirectional gated recurrent unit (BiGRU) (Johnson & Zhang, 2016), and convolutional neural network (CNN) (Deliu et al., 2017). MT-based approaches are hacker asset detection methods that rely on Google Translate (Li et al., 2016; Samtani et al., 2017). These models leverage both English and non-English language labeled data for training and include two state-of-the-art families of hacker asset detec-

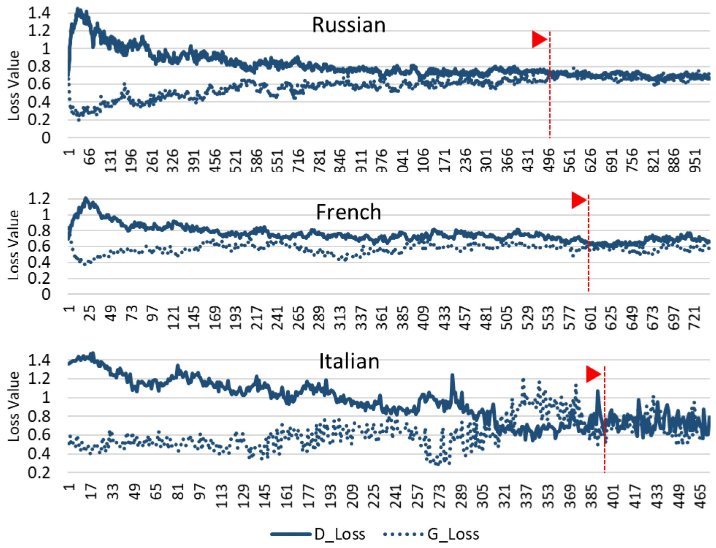


Figure 4. Discriminator and Generator Loss During ADREL's Training for Russian, French, and Italian

tion methods: (1) traditional ML approaches, including naïve Bayes (Samtani et al., 2017), SVM (Deliu et al., 2017), and random forest (Portnoff, 2018), and (2) deep learning methods, including BiLSTM, BiGRU (Grisham et al., 2017) and CNN (Deliu et al., 2017). CLKT alternatives are concerned with how including both English and untranslated non-English text as input affects classification performance. Two recent CLKT models were selected: fully multi-language CNN (FML-CNN) and multi-task learning BiLSTM (MTL-BiLSTM and MTL-BiGRU). FML-CNN combines all English and non-English content into a bag-of-words and inputs it into a three-layer CNN (Deriu et al., 2017). MTL-BiLSTM and MTL-BiGRU utilize multi-task learning to create a shared representation from separate inputs simultaneously (Ebrahimi et al., 2018). We summarize the training data for each benchmark method category in Table 2.

Evaluation Metrics. Consistent with the past security analytics studies, we used three well-established performance metrics to evaluate the performance of all detection methods: accuracy, F_1-score, and area under the ROC curve (AUC) (Deliu et al., 2017; Li et al., 2016). Due to the class imbalance, accuracy alone is not a reliable measure to truly reflect the hacker asset detection performance (Wheelus et al., 2018). It is often recommended to use the F_1-score, which is the harmonic mean of precision and recall (Li et al., 2016). AUC is the most effective measure for class-imbalanced data (Wheelus et al., 2018), which establishes a trade-off between Type I and Type II errors. To help ensure model generalizability, we employed five-fold cross-validation in all experiments, and further assessed the statistical significance of the results by paired *t*-test (Li et al., 2016). The baseline lexicon search does not output probabilities; therefore, AUC is not applicable. The evaluation results for hacker forums and DNMs are summarized in Table 3. The highest scores are highlighted in boldface. As shown in Table 3, CLHAD outperformed all benchmark methods for both Russian and French in terms of accuracy, F_1-score, and AUC with statistically significant margins. Within Russian hacker forums, CLHAD resulted in almost 12% improvement in AUC compared to the second-best performing benchmark method (i.e., 82.33% vs. 70.18% from FML-CNN). Similarly, in French hacker forums, ADREL improved the AUC by approximately 14% (i.e., 85.80% vs. 71.47% from MTL-BiLSTM). Similar results were observed in Russian, French, and Italian DNMs. Results on Italian DNMs are presented in Appendix A. Higher performance in DNMs could be attributed to less generic and more explicit language that is used in product descriptions to maximize profit.

From Table 3, CLHAD's strong performance could be attributed to three factors. First, observing that MT-based methods do not necessarily perform better than monolingual models suggests that mistranslations can negatively affect hacker asset detection performance. Second, CLHAD's strong performance over deep learning monolingual models indicates that generating language-specific representations is insufficient for CLKT purposes. However, transferring salient features from English enhances overall hacker asset detection performance. Third, CLHAD's improved performance over

Table 2. Training Data for Each Category of Benchmark Method

Benchmark Method Category	Training Data Language	Training Data Language Category
Baseline (Lexicon-based)	• Only Russian • Only French • Only Italian	Low-resource non-English
Monolingual		
Machine Translation (MT)-based	• English + Russian • English + French • English + Italian	English (source) + Low resource non-English (target)
CLKT alternatives		
CLHAD (Proposed Method)		

Table 3. Benchmark Evaluation Results in Hacker Forums and DNMs for Russian and French

			Hacker Forums			DNMs		
Method Category	Method		Acc (%)	F_1 (%)	AUC (%)	Acc (%)	F_1 (%)	AUC (%)
Baseline	Lexicon Search	Ru	70.00**	53.06***	N/A	72.61***	53.33***	N/A
		Fr	72.50*	59.26*	N/A	61.43***	58.34**	N/A
Monolingual	NB	Ru	62.82***	62.31***	63.07**	95.00*	88.94*	90.95**
		Fr	66.47**	37.68***	60.38***	72.86*	60.36**	71.83*
	SVM	Ru	68.60**	39.17**	62.83***	94.17*	91.97*	94.02**
		Fr	69.30**	32.66***	60.55***	69.05*	51.85*	66.25**
	RF	Ru	67.39**	37.21***	59.74***	93.41**	90.73**	92.34*
		Fr	70.23**	31.29***	58.40***	68.85*	42.70***	60.18***
	BiGRU	Ru	73.59*	57.12**	69.21**	95.37*	92.81*	96.85*
		Fr	73.08*	54.92***	69.40**	72.86*	52.72**	75.72*
	BiLSTM	Ru	69.43*	57.75*	68.25*	93.45*	89.41*	96.90*
		Fr	74.92*	61.69**	73.54*	66.19**	45.66**	71.88*
	CNN	Ru	68.92*	55.40**	63.75**	93.92*	91.98*	96.31*
		Fr	70.29*	55.74***	73.21*	68.57*	57.07*	70.27**
MT-Based	NB+MT	Ru	62.29***	35.42**	56.32***	95.27***	92.31*	94.16**
		Fr	61.93***	65.57*	62.64***	81.66*	65.78***	76.07***
	SVM+MT	Ru	64.37***	52.45***	62.32***	74.00***	62.62**	79.09*
		Fr	62.28***	55.85***	61.98***	68.33*	38.00**	63.78**
	RF+MT	Ru	64.65**	53.62**	63.79***	83.00*	63.34*	74.39**
		Fr	67.02**	60.82*	66.67**	76.06*	63.61*	74.08*
	BiGRU+MT	Ru	61.60***	63.05*	65.51***	96.31*	91.98*	96.69***
		Fr	61.23***	66.83*	63.52***	77.49*	62.19**	79.77*
	BiLSTM+MT	Ru	63.57**	63.48*	66.32***	93.98***	88.88***	98.02*
		Fr	66.31***	67.28*	68.29***	73.44*	64.15**	78.96*
	CNN+MT	Ru	62.91**	57.03*	63.87**	89.09**	93.12*	94.11**
		Fr	63.33***	66.53*	69.05***	66.88**	67.70**	74.39**
CLKT Alternatives	FML-CNN	Ru	66.11*	58.97**	70.18**	94.36**	91.05**	96.98**
		Fr	75.20*	56.82***	61.27***	68.09*	57.07**	70.31**
	MTL-BiGRU	Ru	64.07*	59.12*	67.25**	94.54*	91.59*	96.05*
		Fr	70.77**	59.54**	68.84***	74.76*	65.68*	77.86*
	MTL-BiLSTM	Ru	69.81*	57.28***	69.98***	95.42**	91.41**	95.58*
		Fr	74.31*	61.30**	71.47**	67.14*	59.68*	71.56*
CLHAD (Proposed Method)		Ru	**79.09**	**70.16**	**82.33**	**97.10**	**95.65**	**98.90**
		Fr	**82.48**	**74.55**	**85.80**	**86.73**	**76.85**	**86.32**

Note: BiGRU: Bidirectional Gated Recurrent Unit; FML-CNN: Fully MultiLingual CNN; MTL-GRU/BiLSTM: Multi-Task Learning GRU/BiLSTM; RF: Random Forest (*p*-values are significant at 0.05:*, 0.01:**, 0.001:***).

prevailing CLKT methods (e.g., FML-CNN, MTL-BiLSTM) indicates that ADREL's adversarial training process leads to language-invariant representations that cannot be attained with simple bag-of-words approaches (e.g., in FML-CNN) or multi-task learning (MTL-BiLSTM). Overall, the consistency of the results shows CLHAD's generalizability across multiple dark web languages. Appendix B further demonstrates CLHAD's generalizability to new languages via empirical analysis of the training size. This empirical analysis shows that knowledge transfer within the same language family (e.g., English and French) requires less training data than languages from different language families (e.g., English and Russian).

Explanation and Detected Hacker Assets Profiling

In addition to the confidence score from CLHAD's output, model explainability is essential for analysts to understand why CLHAD yields specific outputs. Hence, we incorporated a well-established model-agnostic explainability mechanism that further explains CLHAD's output. We then examined the output to obtain a hacker asset profile for each language.

Results Explanation

Consistent with text representation learning literature, we adopted a well-established method called Shapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017). SHAP employs cooperative game theory to quantify the contribution of each word (known as Shapley value) to the final decision made by CLHAD. Three representative samples of hacker assets detected by CLHAD are presented in Table 4. The examples show a spyware program for accessing victims' wallets and social media accounts (Russian), a database injection tool on Windows to recover user logs (French), and a MAC address spoofing tool (Italian). The output explanation column shows the top five contributing words to CLHAD's decision, based on the coefficients of logistic regression. Translations of these words are given in parentheses and are highlighted in the original text based on their contribution (darker color denotes higher contribution).

Words such as "victim," "spy," "user," and "data" contribute to the detection of the Russian spyware as a hacker asset. Similarly, the appearance of words such as "logs" and "inject" in the French database injection tool and words such as "MAC," "address," and "spoof" in the Italian MAC address spoofing tool contribute to detection of these hacker assets by CLHAD. At this stage, a simple lexicon translation can be effectively used by a non-native analyst to assess the results. The incorporated explainability mechanism provides a useful tool to analyze CLHAD's output.

Detected Hacker Assets Profiling

CLHAD's results are also useful to profile hacker assets in non-English platforms. We further examined the hacker assets discovered by CLHAD in our gold-standard dataset via lexicon search to create four groups of assets: hacking tools, hacking services (e.g., DDoS, targeted spying), hacking tutorials, and financial fraud tools (e.g., credit card cloning software). Figure 5 presents the resulting hacker asset profile for dark web platforms by language. The horizontal axis shows the percentage of detected hacker assets in each category.

Three important observations are made from the hacker asset profile shown in Figure 5. First, while all platforms almost equally concentrate on financial fraud as a lucrative business (diamond pattern), English and Russian platforms are mostly focused on hacking tools (horizontal stripes). Second, Russian platforms are different from others in that they heavily focus on hacking services (vertical stripes). An example of hacking services is shown below:

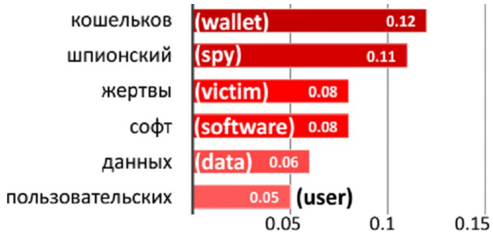
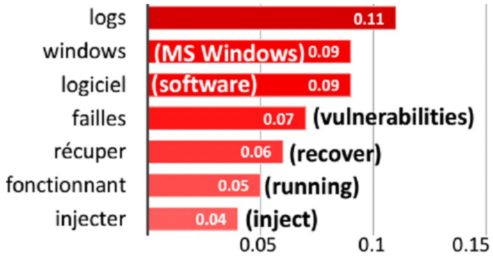
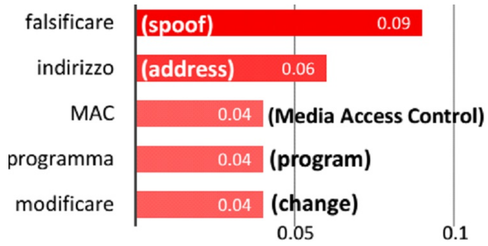
> "Взлом аккаунта Skype. Сообщаете адрес аккаунта, который нужно взломать и оплачиваете сделку …" (translation: "Hacking Skype account. Provide the address of the account you want to hack and pay for the deal ...").

These services often require sophisticated hacking skills and sometimes are indicators of insiders' illegal access to proprietary datasets. Hacking services are not common in English, French, or Italian platforms. This could indicate the concentration of advanced organized hackers in Russia. Finally, while French and Italian platforms highly focus on hacking tutorials, Russian platforms are less focused on tutorials. This may also signify a higher level of hacking skills in Russian platforms as opposed to French and Italian platforms, which may serve lower-skilled hackers or script-kiddies.

Discussion: Contributions to the IS Knowledge Base and Managerial Implications

Guided by the computational design science paradigm, we targeted an emergent cybersecurity application and extensively explored and evaluated viable solutions to design a novel IT artifact suitable for our research inquiry of interest. As a result, several key contributions were made to the IS knowledge base that can help guide future research. We discuss these contributions and their related managerial implications in the following subsections.

Table 4. Explanation of Examples Identified by CHLAD

Original Content	Output Explanation
Предлагаем шпионский софт для получения пользовательских данных к страницам в соцсетях, почты, мессенджеров, кошельков, скриншотов с экрана компьютера жертвы.	
Logiciel fonctionnant sous windows afin de pouvoir traver et injecter des failles SQL afin de pouvoir récuper des logs.	
Con il nostro programma, ora puoi facilmente falsificare l'indirizzo MAC. Con pochi clic, gli utenti saranno in grado di modificare i propri indirizzi MAC.	

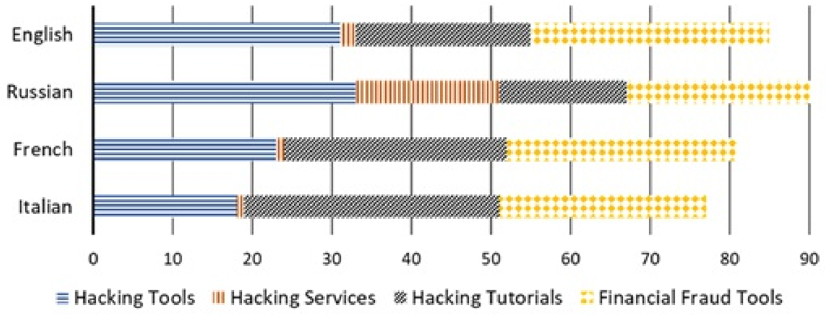


Figure 5. Hacker Asset Profile for Dark Web Platform by Language

Contributions to the IS Knowledge Base

Our study contributes two novel design principles (Gregor & Hevner, 2013) to the IS knowledge base: (1) leveraging multiple languages simultaneously to create comprehensive text representations for downstream text analysis tasks and (2) transferring the expert knowledge (i.e., human-labeled data) to new domains via domain-invariant representations. These design principles can facilitate novel research inquiries within cybersecurity and social media analytics. Within cybersecurity analytics, design principles 1 and 2 offer a more effective mechanism than traditional text processing methods to process unstructured dark web content. Both design principles can also facilitate other cybersecurity analytics tasks, such as identifying key hackers and cybercriminal communities in non-English platforms. Within social media analytics, these design principles can offer language-invariant text representations for content moderation in foreign-language social media platforms. They are also applicable to community question answering (CQA) platforms that aim to harness the collective intelligence of numerous geopolitical regions by allowing the questions and answers to be written in non-English languages by participants globally.

Managerial Implications for Cybersecurity Analytics

CLHAD findings benefit cybersecurity managers at operational and strategic levels. Within operational security management, we believe two findings from CLHAD could benefit information security officers (ISOs) and practitioners in cybersecurity analytics organizations. First, discovering that Russian cybercriminals are more likely to be equipped with sophisticated hacking skills can provide insights into cyber attack attribution (a crucial task in incident response). Accordingly, practitioners of cybersecurity analytics organizations and ISOs in security operation centers (SOCs) can consider automated hacker asset profiling to support attack attribution efforts. For example, they can initially focus on certain geopolitical regions that are more likely to be attributed to advanced nation-state cyber attacks before conducting expensive fully-fledged investigations around the globe. Second, discovering that cybercriminals in Russian, French, Italian, and English dark web platforms almost equally concentrate on financial fraud as a lucrative business, suggests that cybersecurity analytics organizations that protect financial firms need to monitor non-English platforms in addition to only English platforms. As such, automated multilingual hacker asset detection helps prioritize hacker assets based on the security needs of firms and the global cybersecurity landscape. Juxtaposing these findings shows that while security managers may benefit more from focusing on Russian platforms in identifying sophisticated hacking assets, identifying financial hacker assets requires attending to other languages in the dark web as well.

Within strategic security management, CLHAD findings could benefit information systems security managers (ISSMs) and chief information security officers (CISOs) in two areas. First, given that dark web content varies in each geopolitical region and language domain, profiling hacker assets in foreign platforms is useful to improve quarterly and/or annual cybersecurity reports. Such reports have a crucial role in informing resource allocation for mitigation strategies and more effective cybersecurity investments. Second, given that hiring cybersecurity experts who are also native speakers in all dark web foreign languages is expensive and inefficient, the analytics from CLHAD can help identify and tailor the need for hiring analysts with specific language proficiencies in security firms. For instance, based on our findings, a large security firm focusing on nation-state hacker assets can benefit from employing analysts with Russian language proficiency. Analysts with such language proficiency could also benefit from the explanations obtained from CLHAD's automated detection and hacker asset profiling.

Conclusion and Future Directions

Detecting hacker assets in massive volumes of dark web content is crucial to gain reconnaissance on adversaries' arsenals. Manual hacker asset detection is *ad hoc*, labor-intensive, costly, not scalable, and time-consuming. There has been a significant push in cybersecurity analytics to develop automated ML-based approaches. However, ML approaches often require analysts to provide labeled data during the model training phase. While providing human-labeled data is more feasible for English content, non-English dark web platforms suffer from a lack of labeled data due to the unfamiliarity of the analysts with foreign languages or issues with MT in the cybersecurity domain. The rapid rise of nation-specific dark web platforms calls for novel multilingual cybersecurity analytics. In this study, we adopted the computational design science paradigm to develop a novel security analytics framework (CLHAD) that enables the explainable detection of hacker assets in non-English dark web platforms using deep cross-lingual knowledge transfer. Our approach employs a novel adversarial learning procedure to capture the language-invariant representations from English and foreign platforms and automatically detect hacker assets in non-English platforms. Through rigorous benchmark evaluations on Russian, French, and Italian hacker forums and dark net markets, we demonstrated that our method significantly improves hacker asset detection across multiple foreign languages.

Future research can build upon CLHAD by integrating its results into social network analysis to identify key hacker communities, inform resource allocation, and help reduce the overall cost of global cyber-crime.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under the grants CNS-1936370 (SaTC CORE), DGE-1921485 (SFS), CNS-1850352 (CRII SaTC), DGE-1303362 (SFS), and SES-1303362 (SBE SaTC TTP). Yidong Chai was designated as the corresponding author.

References

Abdalla, M., & Hirst, G. (2017). Cross-lingual sentiment analysis without (good) translation. In *International Joint Conference on Natural Language Processing*, Taiwan, 506–515.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (pp. 214–223). JMLR.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828.

Benjamin, V., & Chen, H. (2015). Developing understanding of hacker language through the use of lexical semantics. In *Proceedings of the 13th IEEE International Conference on Intelligence and Security Informatics* (pp. 79–84). IEEE.

Benjamin, V., & Chen, H. (2016). Identifying language groups within multilingual cybercriminal forums. In *IEEE International Conference on Intelligence and Security Informatics* (pp. 205–207). IEEE.

Benjamin, V., Valacich, J., & Chen, H. (2019). DICE-E: A framework for conducting darknet identification, collection, evaluation with ethics. *MIS Quarterly*, 43(1), 1–22.

Benjamin, V., Zhang, B., Nunamaker Jr., J. F., & Chen, H. (2016). Examining hacker participation length in cybercriminal internet-relay-chat communities. *Journal of Management Information Systems*, 33(2), 482–510.

Cao, Q., & Xiong, D. (2018). Encoding gated translation memory into neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3042–3047). Association for Computational Linguistics.

Chen, H. (2012). *Dark web: Exploring and data mining the dark side of the web*. Springer.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188.

Deliu, I., Leichter, C., & Franke, K. (2017). Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks. In *Proceedings of the 2017 IEEE International Conference on Big Data* (pp. 3648–3656). IEEE.

Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., & Jaggi, M. (2017). Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In Proceedings of the 26th *International Conference on World Wide Web*, 1045–1052.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.

Dittus, M., Wright, J., & Graham, M. (2018). Platform criminalism: The "last-mile" geography of the darknet market supply chain. In *Proceedings of the 2018 World Wide Web Conference* (pp. 277–286). International World Wide Web Conference Committee.

Dong, X., & de Melo, G. (2018). Cross-lingual propagation for deep sentiment analysis. In *Proceedings of the 21nd AAAI Conference on Artificial Intelligence* (pp. 5771–5778). AAAI Press.

Duek, S., & Markovitch, S. (2018). Automatic generation of language-independent features for cross-lingual classification. Available through *arVix* (https://arxiv.org/abs/1802.04028). Association for the Advancement of Artificial Intelligence.

Duong, L., Kanayama, H., Ma, T., Bird, S., & Cohn, T. (2016). Learning crosslingual word embeddings without bilingual corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1285–1295). Association for Computational Linguistics.

Ebrahimi, M., Nunamaker Jr., J. F., & Hsinchun, C. (2020). Semi-supervised cyber threat identification in dark net markets: A transductive and deep learning approach. *Journal of Management Information Systems*, 37(3), 694–722.

Ebrahimi, M., Surdeanu, M., Samtani, S., & Chen, H. (2018). Detecting cyber threats in non-English dark net markets: A cross-lingual transfer learning approach. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics* (pp. 85–90). IEEE.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1), 2096–2030.

Goldberg, Y. (2017). *Neural network methods for natural language processing*. Morgan-Claypool Publishers.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning*. MIT Press.

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337–355.

Grisham, J., Samtani, S., Patton, M., & Chen, H. (2017). Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence. In *Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics* (pp. 13–18). IEEE.

Hui, K.-L., Kim, S. H., & Wang, Q.-H. (2017). Cybercrime deterrence and international legislation: Evidence from distributed denial of service attacks. *MIS Quarterly*, 41(2), 497–523.

Jensen, M. L., Dinger, M., Wright, R. T., & Thatcher, J. B. (2017). Training to mitigate phishing attacks using mindfulness techniques. *Journal of Management Information Systems*, 34(2), 597–626.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5, 339–351.

Johnson, R., & Zhang, T. (2016). Supervised and semi-supervised text categorization using LSTM for region embeddings. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning* (pp. 526–534).. JMLR.org.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). *Exploring the limits of language modeling*. Google Inc. (www.ai.google/research/pubs/pub45446).

Li, N., Zhai, S., Zhang, Z., & Liu, B. (2017). Structural correspondence learning for cross-lingual sentiment classification with one-to-many mappings. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence* (pp. 3490–3496). AAAI Press.

Li, W., Chen, H., & Nunamaker Jr., J. F. (2016). Identifying and profiling key sellers in cyber carding community: AZSecure text mining system. *Journal of Management Information Systems*, 33(4), 1059–1086.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 21st Conference on Neural Information Processing Systems* (https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf).

Marin, E., Diab, A., & Shakarian, P. (2016). Product offerings in malicious hacker markets. In *Proceedings of the IEEE Conference on Intelligence and Security Informatics* (pp. 187–189). IEEE.

Morgan, S. (2017). 2017 cybercrime report: Cybercrime will cost the world $6 trillion annually by 2021. Cybersecurity Ventures, Herjavec Group.

Ning, Y., Wu, Z., Li, R., Jia, J., Xu, M., Meng, H., & Cai, L. (2017). Learning cross-lingual knowledge with multilingual BLSTM for emphasis detection with limited training data. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5615–5619). IEEE.

Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V., Robertson, J., Shakarian, J., Thart, A., & Shakarian, P. (2016). Darknet and deepnet mining for proactive cybersecurity threat intelligence. In *Proceedings of the IEEE Conference on Intelligence and Security Informatics* (pp. 7–12). IEEE.

Pastrana, S., Thomas, D. R., Hutchings, A., & Clayton, R. (2018). CrimeBB: Enabling cybercrime research on underground forums at scale. In *Proceedings of the 2018 World Wide Web Conference* (pp. 1845–1854). International World Wide Web Conference Committee.

Portnoff, R. (2018). *The dark net: De-anonymization, classification and analysis*. Unpublished Ph.D. Thesis, EECS Department, University of California, Berkeley. (www.eecs. berkeley.edu/Pubs/TechRpts/2018/EECS-2018-5.html).

Queiroz, A. L., Mckeever, S., & Keegan, B. (2019). Detecting hacker threats: Performance of word and sentence embedding models in identifying hacker communications. In *Proceedings of the 27th AIAI Irish Conference on Artificial Intelligence and Computer Science* (pp. 116–127).

Rai, A. (2017). Editor's comments: Diversity of design science research. *MIS Quarterly*, 41(1), iii–xviii.

Rasooli, M. S., Farra, N., Radeva, A., Yu, T., & McKeown, K. (2018). Cross-lingual sentiment transfer with limited resources. *Machine Translation*, 32(1), 143–165.

Samtani, S., Chinn, R., Chen, H., & Nunamaker Jr., J. F. (2017). Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence. *Journal of Management Information Systems*, 34(4), 1023–1053.

Schäfer, M., Fuchs, M., Strohmeier, M., Engel, M., Liechti, M., & Lenders, V. (2019). BlackWidow: Monitoring the dark web for cyber security information. In *Proceedings of the International Conference on Cyber Conflict* (pp. 1–21). IEEE.

Shore, J., Baek, J., & Dellarocas, C. (2018). Network structure and patterns of information diversity on Twitter. *MIS Quarterly*, 42(3), 849–872.

Spataro, J. (2021). Tactical linguistics: Language analysis in cyber threat intelligence. SANS Institute, January 15 (https://www.sans.org/reading-room/whitepapers/threatintelligence/paper/40075).

Temizkan, O., Park, S., & Saydam, C. (2017). Software diversity for improved network security: Optimal distribution of software-based shared vulnerabilities. *Information Systems Research*, 28(4), 828–849.

Tian, J., Lan, M., Wu, Y., Wang, J., Qiu, L., Li, S., Jun, L., & Si, L. (2018). An adversarial joint learning model for low-resource language semantic textual similarity. In G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury (Eds.), *Advances in Information Retrieval* (pp. 89–101). Springer.

Tolido, R., Linden, G. van der, Delabarre, L., Theisler, J., Khemka, Y., Thieullent, A.-L., Frank, A., Buvat, J., & Cherian, S. (2019). Reinventing cybersecurity with artificial intelligence: The new frontier in digital security. Capgemini Research Institute (www.capgemini.com).

Vance, A., Jenkins, J. L., Anderson, B. B., Bjornn, D. K., & Kirwan, C. B. (2018). Tuning out security warnings: A longitudinal examination of habituation through fMRI, eye tracking, and field experiments. *MIS Quarterly*, 42(2), 355–380.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1), 1–40.

Wheelus, C., Bou-Harb, E., & Zhu, X. (2018). Tackling class imbalance in cyber security datasets. In *Proceedings of the 2018 IEEE International Conference on Information Reuse and Integration* (pp. 229–232). IEEE.

Xu, R., & Yang, Y. (2017). Cross-lingual distillation for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 1415–1425). Association for Computational Linguistics.

Yang, S., Hsu, C., Sarker, S., & Lee, A. S. (2017). Enabling effective operational risk management in a financial institution: An action research study. *Journal of Management Information Systems*, 34(3), 727–753.

Yang, X., McCreadie, R., Macdonald, C., & Ounis, I. (2017). Transfer learning for multi-language twitter election classification. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 341–348). ACM Press.

Yin, H. S., Langenheldt, K., Harlev, M., Mukkamala, R. R., & Vatrapu, R. (2019). Regulating cryptocurrencies: A supervised machine learning approach to de-anonymizing the bitcoin blockchain. *Journal of Management Information Systems*, 36(1), 37–73.

Yuan, K., Lu, H., Liao, X., & Wang, X. (2018). Reading thieves' cant: Automatically identifying and understanding dark jargons from cybercrime marketplaces. In *Proceedings of the 27th USENIX Security Symposium* (pp. 1027–1041). USENIX.

Yue, W. T., Wang, Q., & Hui, K.-L. (2019). See no evil, hear no evil? Dissecting the impact of online hacker forums. *MIS Quarterly*, 43(1), 73–95.

Zhou, X., Wan, X., & Xiao, J. (2016). Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1403–1412). Association for Computational Linguistics.

About the Authors

Mohammadreza (Reza) Ebrahimi is an assistant professor in the School of Information Systems and Management at the University of South Florida. Reza received his Ph.D. in Management Information Systems from the University of Arizona, where he was a research associate at the Artificial Intelligence (AI) Lab in 2021. He received his master's degree in Computer Science from Concordia University, Canada, in 2016. Reza's dissertation on AI-enabled cybersecurity analytics received the ICIS ACM SIGMIS best doctoral dissertation award in 2021. His current research is focused on statistical and adversarial machine learning theories for AI-enabled secure and trustworthy cyberspace. Reza has published 25 articles in peer-reviewed journals, conferences, and workshops, including *Journal of Management Information*, *IEEE TPAMI, Applied Artificial Intelligence, Digital Forensics*, *IEEE S&PW, AAAI, IEEE ICDM,* and *IEEE ISI.* He has served as a program chair and program committee member for the IEEE ICDM Workshop on Machine Learning for Cybersecurity (MLC) and the IEEE S&P Workshop on Deep Learning and Security (DLS). He has contributed to several projects supported by the National Science Foundation. Reza is a member of the IEEE, ACM, AAAI, and AIS.

Yidong Chai received his bachelor's degree in information systems from Beijing Institute of Technology, and his Ph.D. degree in Management Information Systems from Tsinghua University. He is currently a professor at the Hefei University of Technology. His research fields include machine learning, signal processing, and natural language processing. His work has appeared in journals including *Knowledge-Based Systems* and *Applied Soft Computing*, as well as conferences and workshops including IEEE S&P, INFORMS Workshop on Data Science, Workshop on Information Technology Systems, International Conference on Smart Health, and International Conference on Information Systems.

Sagar Samtani is an assistant professor and Grant Thornton Scholar in the Department of Operations and Decision Technologies at the Kelley School of Business at Indiana University. Sagar graduated with his Ph.D. in Management Information Systems from the University of Arizona's Artificial Intelligence Lab in May 2018, where he served as a Scholarship-for-Service Fellow. Sagar's AI for cybersecurity and dark web analytics research initiatives have received funding from the National Science Foundation CRII, CICI, and SaTC-EDU programs. Sagar has published over 40 articles in peer reviewed journals and conferences including *MIS Quarterly*, *Journal of MIS*, *IEEE Intelligent Systems*, *ACM Transactions on Privacy and Security*, and others. He is currently an associate editor for *Information and Management* and has served as a guest editor for *IEEE Transactions on Dependable and Secure Computing* and *ACM Transactions on MIS*. His research has received multiple awards and significant media coverage and citations from outlets such as the Miami Herald, Fox News, and Science.

Hsinchun Chen is Regents Professor and Thomas R. Brown Chair in Management and Technology in the Management Information Systems Department at the Eller College of Management, University of Arizona. He received his Ph.D. in Information Systems from New York University. He is the author/editor of 20 books, 300 journal articles, and 200 refereed conference articles covering digital library, data/text/web mining, business analytics, security informatics, and health informatics. He founded the Artificial Intelligence Lab at The University of Arizona in 1989, which has received $50M+ research funding from the NSF, National Institutes of Health, National Library of Medicine, Department of Defense, Department of Justice, Central Intelligence Agency, Department of Homeland Security, and other agencies (100+ grants, 50+ from NSF). His COPLINK/i2 system for security analytics was commercialized in 2000 and acquired by IBM as its leading government analytics product in 2011. The COPLINK/i2 system is used in 5,000+ law enforcement jurisdictions and intelligence agencies in the U.S. and Europe, making a significant contribution to public safety worldwide.

Appendix A

Benchmark Evaluation of (CLHAD) on Italian DNMs

We present the performance of CLHAD against baseline, monolingual, MT-based, and CLKT methods on Italian dark web platforms in Table A1. The highest performance is shown in bold-face.

Table A1. Benchmark **Evaluation** Results in Italian DNMs

Method Category	Method	Accuracy	F_1-score	AUC
Baseline	Lexicon Search	72.64**	61.45***	N/A
Monolingual	NB	71.03***	59.44***	73.07***
	SVM	78.90*	69.17*	78.30**
	RF	79.29*	69.87*	80.28**
	BiGRU	81.75*	69.75*	84.12**
	BiLSTM	80.83*	69.11*	84.71*
	CNN	81.75*	67.78*	85.16**
MT-Based	NB+MT	66.15**	49.30*	64.23**
	SVM+MT	79.77*	54.15***	67.44***
	RF+MT	71.51**	41.11**	62.47***
	BiGRU+MT	85.34**	72.23***	87.94**
	BiLSTM+MT	84.27**	71.92***	87.44**
	CNN+MT	73.49***	73.02*	79.53*
CLKT Alternatives	FML-CNN	81.75*	67.78*	85.88**
	MTL-BiGRU	83.44*	72.39*	85.98**
	MTL-BiLSTM	82.75*	71.35*	85.76*
CLHAD (Proposed Method)		**88.16**	**79.59**	**91.68**

CLHAD outperformed the benchmark methods in AUC, F_1-score, and accuracy with statistically significant margins. Compared to the second-best methods for each metric, CLHAD improved hacker asset detection in Italian platforms by almost 7% in F_1-score (vs. MTL-BiGRU in CLKT category), and by 4% in AUC (vs. BiGRU+MT in MT-based category).

Appendix B

Empirical Analysis of the Training Size in Target Languages

To empirically examine the size of training set needed for knowledge transfer to the target language, we started with 20% of the original training size and gradually increased the size of our training set. We measured the performance of the proposed method at each step. Figure B1 depicts the results for all performance measures while varying different training sizes (0.2 denotes an 80% reduction in the training size, and 1 denotes the original training size). As shown in Figure B1, to achieve an AUC above 80%, the model needs to be trained on at least 1,817 Russian and 570 French documents. For Italian, this requirement is even lower (261 documents to achieve 80% AUC). The minimum AUC is shown by an arrow. The corresponding minimum training size is shown by a dashed line.

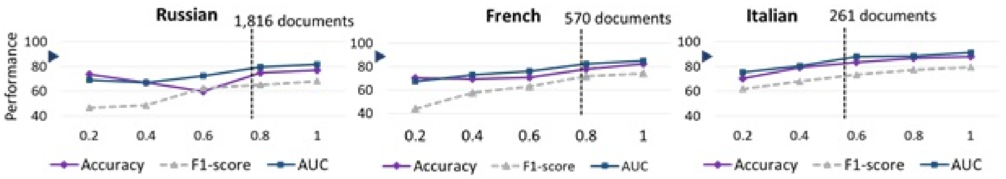


Figure B1. Empirical Analysis of the Training Size in Target Languages

The empirical results could suggest that since French and Italian are more similar to English (both are close descendants of the Indoeuropean family) than Russian, their minimum required size of training data is far less than that of Russian (descended from the Slavic family). Given the minimum required number of training documents in French and Italian, it is expected that achieving similar performance in a new language with similar characteristics to English (i.e., the source language) would need no more than several hundred labeled training documents. However, learning a new target language that is farther from the English family (e.g., Russian) would need a larger training set of one or two thousand documents. The empirical results align with the intuition that knowledge transfer from English to similar target languages (from the same family) requires less training data than languages that are from different language families.