Detecting Cyber Threats in Non-English Hacker Forums: An Adversarial Cross-Lingual Knowledge Transfer Approach

Mohammadreza Ebrahimi
Artificial Intelligence Lab
Department of Management Information Systems
University of Arizona
Tucson, Arizona, USA
ebrahimi@email.arizona.edu

Sagar Samtani
Department of Operations and Decision Technologies
Indiana University
Bloomington, Indiana, USA
sagarsamtani96@gmail.com

Yidong Chai*
Department of management Science and Engineering
Tsinghua University
Beijing, China
chaiyd14@mails.tsinghua.edu.cn

Hsinchun Chen
Artificial Intelligence Lab
Department of Management Information Systems
University of Arizona
Tucson, Arizona, USA
hchen@eller.arizona.edu

***Abstract*—The regularity of devastating cyber-attacks has made cybersecurity a grand societal challenge. Many cybersecurity professionals are closely examining the international Dark Web to proactively pinpoint potential cyber threats. Despite its potential, the Dark Web contains hundreds of thousands of non-English posts. While machine translation is the prevailing approach to process non-English text, applying MT on hacker forum text results in mistranslations. In this study, we draw upon Long-Short Term Memory (LSTM), Cross-Lingual Knowledge Transfer (CLKT), and Generative Adversarial Networks (GANs) principles to design a novel Adversarial CLKT (A-CLKT) approach. A-CLKT operates on untranslated text to retain the original semantics of the language and leverages the collective knowledge about cyber threats across languages to create a language invariant representation without any manual feature engineering or external resources. Three experiments demonstrate how A-CLKT outperforms state-of-the-art machine learning, deep learning, and CLKT algorithms in identifying cyber-threats in French and Russian forums.**

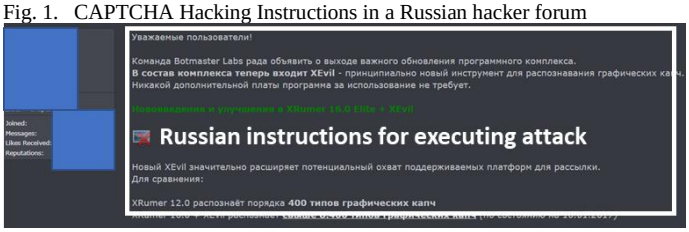
Keywords— adversarial learning, generative adversarial networks, hacker forums, cross-lingual knowledge transfer, long short-term memory

I. INTRODUCTION

The regularity and disturbing frequency of devastating cyber-attacks has made cybersecurity a grand societal challenge. Cyber-analysts in numerous public, private, and academic organizations are increasingly relying on methods to automatically sift through large quantities of cybersecurity relevant data (e.g., log files) to detect potential cyber threats. However, the continuing growth of cyber-attacks indicates that analyzing attacks after they occur cannot keep up with the ever-growing threat landscape. Consequently, innovative approaches to proactively identifying cyber threats are critically needed.

Numerous cybersecurity professionals are turning to the Dark Web to proactively identify cyber threats. The Dark Web is a dark covert side of the web that allows hackers to share, sell, and discuss hacking tools, knowledge, and other cyber threats across multiple geopolitical regions such as the US, Russia, France, and others [1]. The Dark Web comprises four major platforms: hacker forums, DarkNet Marketplaces (DNMs), carding shops, and internet-relay chat (IRC) [2]. Among the four, hacker forums are the largest, often containing hundreds of thousands to millions of cyber threats. While English forums are the most prevalent, Russian and French forums contain a significant quantity of cyber threats such as credit card stealing tools (e.g., skimmers), stolen goods, and others. Fig. 1 illustrates an example of a hacker providing a tool to hack CAPTCHAs in a Russian hacker forum.

Fig. 1. CAPTCHA Hacking Instructions in a Russian hacker forum



Despite their value, the multi-lingual and large volume of non-English forum content poses a significant challenge to cybersecurity analysts and researchers aiming to identify cyber threats. Machine translation (MT) (e.g., Google Translate) is the prevailing approach to process non-English text for subsequent input into a machine learning algorithm. However, applying MT on jargon-laden hacker forum text results in numerous mistranslations that affect cyber threat detection performance. Additionally, past studies process each language separately, rather than leveraging knowledge across the community to detect cyber threats. These limitations require a novel approach that operates on untranslated text and holistically leverages knowledge across languages to detect cyber threats.

In this study, we draw upon Long-Short Term Memory (LSTM), Cross-Lingual Knowledge Transfer (CLKT), and Generative Adversarial Networks (GANs) principles to design a novel Adversarial CLKT (A-CLKT) approach. A-CLKT operates on untranslated text to retain the original semantics of the language. Through an innovative GAN architecture, A-CLKT leverages the collective knowledge about cyber threats across languages to create a language invariant representation (i.e., embedding, feature vector) without any manual feature engineering or external resources. Three experiments demonstrate how A-CLKT outperforms common machine

* Corresponding author

learning, deep learning, and CLKT algorithms in identifying cyber-threats in French and Russian forums. To facilitate scientific reproducibility, we release our code and data through a public GitHub repository.

The remainder of this paper is organized as follows. First, we review related research in hacker forum analysis, CLKT, LSTMs, and GANs. Second, we summarize key research gaps and pose several research questions. Third, we present our proposed research framework and detail its constituent components. Subsequently, we present key experiment results and discuss their implications. Finally, we conclude this research and identify promising future research directions.

II. Literature Review

Four areas of literature are examined to ground this research. First, we review studies on hacker forums to identify their content, data characteristics, and prevailing analytics. Second, we review CLKT to identify the principles of knowledge transfer across languages. Third, we review LSTMs to identify how the state-of-the-art deep learning approach for text operates. Finally, we review GANs as a mechanism for using LSTM's learned representations for CLKT.

A. Hacker Forum Analysis

As indicated in the introduction, hacker forums play a valuable role in the international Dark Web ecosystem by providing millions of hackers the ability to share and discuss cyber threat information and content. Over the past decade, numerous practitioners and scholars have found significant cyber threat content [2]–[7]. Examples include:

- **Hacking tools:** software designed to circumvent security controls and illicitly manipulate technologies (e.g., ransomware, spyware, etc.)
- **Malicious tutorials:** guides instructing hackers on selected tasks (e.g., how to steal cryptocurrency, etc.)
- **Stolen digital goods:** accounts, credentials, and other content attained from hacking targeted victims
- **Credit card fraud:** content to conduct credit card crimes (e.g., skimming, cloning, etc.)

The complexity of this content has resulted in significant natural, non-natural, and jargon-laden text. The prevailing approach to processing non-English forum content is MT to convert all content to English. Translated content has served as input to machine learning algorithms such as recurrent neural networks (RNN) to identify mobile malware [8], maximum entropy and recursive neural networks to detect and rate carding threats [9], and support vector machine (SVM) to categorize hacking tools into their programming languages [10][11].

Despite their convenience, MT-based approaches have three key drawbacks. First, they omit the original, language-specific semantics. Second, MT services are trained on general corpora. Therefore, they often miss hacker specific jargon. Finally, past studies use monolingual models (i.e., separate models) for each language. This does not leverage the language-specific knowledge in each language. Taken together, these issues result in mistranslations and incomplete language representations that deteriorate model performance. In light of these drawbacks, we review CLKT as a possible approach to transfer language-specific knowledge across languages without MT.

B. Cross-Lingual Knowledge Transfer (CLKT)

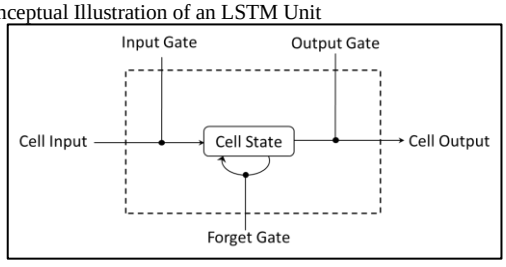
CLKT is a form of transfer learning that aims to learn and transfer the knowledge within a high-resource, source language with significant training data to a low-resource (i.e., limited training data), target language [12]. CLKT is facilitated by learning a representation from the high resource source language and transferring it to the low-resource target language. Three approaches exist to learn and transfer representations: parallel corpora, MT, and pre-trained embeddings. Parallel corpora rely on the alignment of words and sentences across language resources to facilitate knowledge transfer. MT converts the source language to the target. Finally, pre-trained embeddings (i.e., feature vectors) are created by training deep learning algorithms on general-purpose corpora.

Despite their widespread use across critical natural language processing (NLP) tasks, the uniqueness and lack of accessible ground truth hacker forum datasets hinders the direct use of these approaches to facilitating CLKT. Building parallel corpora requires significant manual effort, domain expertise, and manual feature engineering to carefully align words and sentences [13]. MT suffers from the same limitations as discussed earlier. Finally, pre-trained embeddings are developed from general-purpose corpora that do not contain hacker specific semantics and jargon [14]. As a result, conducting CLKT across hacker forum languages requires generating domain-specific embeddings. The prevailing deep learning architecture for this task is LSTM.

C. Long Short-Term Memory (LSTM)

LSTM belongs to an emerging class of deep learning architectures known as recurrent neural networks (RNN), which are designed to analyze sequential data by considering long term dependencies in the input sequence. Such models exhibit strong performance in automatically learning an embedding from a sequence (e.g., words in written language). As a result, LSTM has been widely adopted for language modeling tasks. At a high level, an LSTM cell encompasses non-linear activation functions along with input, output, and forget gates. Fig. 2 illustrates how LSTM uses these components.

Fig. 2. Conceptual Illustration of an LSTM Unit



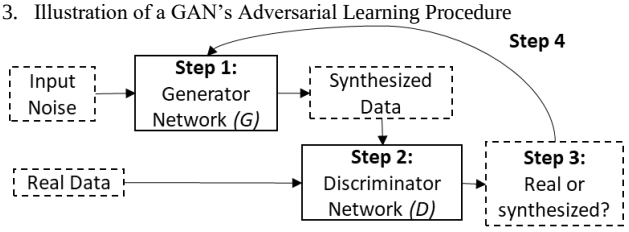
The input gate controls how often a new input token would affect the current cell state. The forget gate determines for how long the cell maintains the current cell state. Finally, the output gate adjusts the effect of the current cell state on the final output of the LSTM cell. The current state of the LSTM cell is obtained as a function of the cell input, input gate, and forget gate. The cell output is attained as a function of the cell state and the output

gate. LSTM offers a viable approach to automatically extract high-quality language-specific representations. However, these representations are not language-invariant, and thus, are less transferable [15]. Therefore, an additional mechanism to transfer the learned representation to other languages is required. One such approach is GAN.

D. Generative Adversarial Networks (GANs)

GAN is a deep learning-based approach that employs an adversarial learning procedure [16]. Adversarial learning is a paradigm within machine learning that has two algorithms compete in a zero-sum game. Within a GAN (Fig. 3), a generator (*G*) network uses input noise to create synthesized data. A discriminator (*D*) aims to distinguish between *G*'s synthesized data from the real data.

Fig. 3. Illustration of a GAN's Adversarial Learning Procedure



GAN's adversarial learning procedure has four steps:

- **Step 1:** Input noise (usually drawn from a uniform distribution) is used by *G* to generate synthesized data.
- **Step 2:** *D* receives the real data and synthesized data from *G* as input and aims to discern between the two.
- **Step 3:** *D*'s prediction is compared with the ground truth with a loss function (e.g., logistic). The errors are backpropagated through *G* to update weights.
- **Step 4:** Steps 1-3 repeat until generator creates data *D* cannot distinguish from the original (i.e., equilibrium)

GAN's adversarial learning strategy is a promising mechanism for transferring an LSTM's learned representation to another language without using external resources. However, how to configure the generator and discriminator such that it can support CLKT is unknown.

III. Research Gaps and Questions

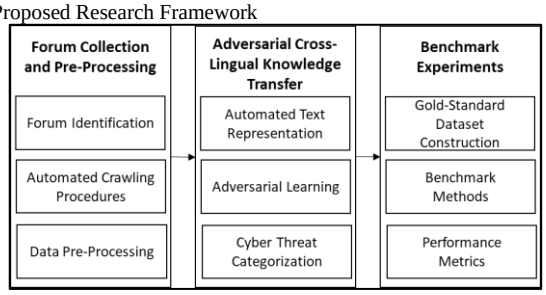
Several research gaps were identified. First, prior hacker forum research uses MT to convert language into English. However, doing so can omit valuable semantics from the original language and result in a deterioration in model performance. Second, while CLKT can transfer knowledge across languages, they often require external resources that are unavailable. Finally, while GAN can potentially transfer an LSTM's learned representations of hacker forum text across languages to enhance cyber detection, how to configure the adversarial learning process accordingly is unclear. Based on these gaps, we propose the following research questions for study:

- How can the adversarial learning procedure within a GAN be extended to account for multiple languages to create a language invariant representation?
- How can an adversarial learning-based CLKT approach that does not require external resources be developed?
- How does an adversarial learning-based CLKT approach perform against prevailing machine learning, deep learning, and CLKT approaches?

IV. Research Design and Testbed

To address the proposed research questions, we designed a novel research framework (Fig. 4) with three major components: (1) Data Collection and Pre-Processing, (2) the proposed A-CLKT, and (3) Benchmark Experiments. Details of each component are presented in the following sub-sections.

Fig. 4. Proposed Research Framework



A. Data Collection and Pre-Processing

Four large-scale and long-running international hacker forums were identified and collected for this study. These forums were identified through three mechanisms. First, we consulted with cybersecurity experts and researchers well-versed in Dark Web analytics and the underground economy. Second, these platforms are well-known within the Dark Web ecosystem as containing a significant quantity of malicious cyber-threats. Third, these forums are highly-ranked in well-known online Dark Web directories, such as Dark Web News.

Following identification, we designed a custom web crawler to collect all forum content. The web crawler was routed through the TOR network to obfuscate our identity. The crawler employs a breadth-first search (BFS) strategy to automatically traverse the forum and parse posts into a database. Table I summarizes each forum's number of posts, authors, and date range. To protect ourselves from hackers within these communities, we denote each forum with a unique identification number.

TABLE I. Summary of Collected Hacker Forums

Language	Forum Name	# of Postings	# of Authors	Date Range
English	b***v	183,354	22,928	2002 – 2018
Russian	a***t	91,667	29,247	2002 – 2018
French	b***k	64,800	9,672	2008 – 2019
	h***s	7,284	1,080	2010 – 2019
Total:	-	**339,821**	**62,927**	**2002 – 2019**

The collection has one English forum, one Russian forum, and two French forums. The posts within these forums were made by 62,927 authors over a 17-year time frame. Forum names are anonymized by asterisks.

Since data pre-processing is critical for algorithm performance, we execute a series of data cleansing tasks. We follow the steps proposed by past multi-lingual CLKT studies [17]. First, we lower-case text to ensure that different cases of the same word are processed identically. Second, we tokenize

the content to split the text into individual units and remove the stop-words. Finally, we unify all tokens to UTF-8 across training and evaluation datasets before constructing trainable word embeddings for each token.

B. Adversarial Cross-Lingual Knowledge Transfer (A-CLKT)

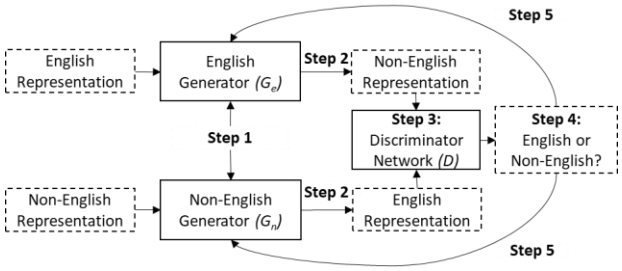
English is considered in our design as the high-resource, source language. The non-English A-CLKT has three major phases: (1) automated text representation, (2) learning a language invariant representation, and (3) cyber threat detection. Each is summarized below.

1) Automated Text Representation: Phase 1 allocates an LSTM for each language to automatically create an embedding of hacker forum text. LSTM is a suitable choice as it was designed for text, performs well on multi-lingual text, and can automatically learn embeddings.

2) Learning a Language Invariant Representation

We devise a novel adversarial learning strategy to operate on the English and non-English representations. Specifically, we formulate the GAN to have two generators and one discriminator. Each generator is assigned to either the English or non-English representation and generate a representation in the *opposite* language. The discriminator aims to distinguish between the generated English and non-English representations. The error between the discriminator's prediction and ground truth is backpropagated to the generators to update weights. This process continues until the discriminator minimizes the error. Fig. 5 illustrates the proposed procedure.

Fig. 5. Illustration of A-CLKT's Adversarial Learning Procedure



- **Step 1:** Each generator is assigned either the English or non-English representation generated from the LSTMs.
- **Step 2:** Each generator generates a representation in the opposite language.
- **Step 3:** D aims to discern between the generated and true English and non-English text.
- **Step 4:** D's prediction is compared with the ground truth with a logistic loss function. The errors are backpropagated to the generators to update weights.
- **Step 5:** Steps 1-4 are repeated until the generators and discriminator reach equilibrium (i.e., generators create English and non-English data that *D* cannot distinguish)

The process described above has several key benefits. First, it operates upon the untranslated hacker forum text. Therefore, it retains the original semantics of the language. Second, it does not require any external resources (e.g., parallel corpora) that are often unavailable for the Dark Web. Finally, the entire process does not require any manual feature engineering. Consequently, it is ideal for rapidly evolving multilingual Dark Web text.

3) Cyber Threat Detection

Phase 3 receives the language invariant representation from the GAN and classifies it as a threat or non-threat. While any classifier can be adopted for this task, the A-CLKT relies on a BiLSTM to conduct the classification. BiLSTM is an extension of the standard LSTM that uses both backward and forward procedures to more comprehensively capture the long and short term dependencies that may occur within a text input.

C. Benchmark Experiments

This section presents the datasets, experiments, benchmark algorithms, and metrics used to evaluate the proposed A-CLKT.

1) Gold-Standard Dataset

Conducting benchmark experiments requires training and evaluating all proposed algorithms on a labeled set of ground-truth (i.e., gold-standard) data. Given the lack of such publicly accessible datasets, we used stratified sampling to extract posts from our collection (Table I). We then assembled a panel of two Russian, two French, and two English cybersecurity experts. Each panelist was assigned to the language of their knowledge and instructed to label individually if a post was threat or non-threat based on its content and their expertise.

After all posts were annotated, we computed the Cohen's kappa coefficient to identify the level of agreement between panelists. For the first round of annotation, the kappa values were 94.48% for English posts, 98.11% for Russian, and 97.01% for French. Additional meetings were held between annotators to resolve differences. After the second round of annotation, the panelists agreed on more than 99% of posts. The unresolved posts were omitted. Table II summarizes the number of labeled cyber-threats and non-cyber threats in each language.

TABLE II. SUMMARY OF GOLD-STANDARD DATASETS USED FOR BENCHMARK EXPERIMENTS

Language	# of Cyber Threats	# of Non-cyber Threats	Total
English	326	1,124	1,450
Russian	83	922	1,005
French	38	464	502
Total:	447	2,510	2,957

Overall, the gold-standard dataset included 1,450 English posts (326 threats, 1,124 non-threats), 1,005 Russian posts (83 threats, 922 non-threats), and 502 French posts (38 threats, 464 non-threats). Since the total number of English posts exceeds the quantity of either the Russian or French, we denote English as the high-resource source language for A-CLKT. Russian or French serve as the low-resource target languages.

2) Experiment 1: A-CLKT vs Machine Translation-based Approaches

Our literature review indicated that the conventional approach to processing non-English hacker forum content is using MT. Therefore, experiment 1 benchmarks A-CLKT's performance against algorithms that use machine-translated content as input. We selected two categories of baseline methods: classical machine learning and deep learning. The former includes naïve Bayes (NB), SVM, random forest (RF),

and k-nearest neighbor (k-NN). This selection represents the prevailing algorithms including decision tree-based (RF), probabilistic (NB), geometric (SVM), and distance-based (k-NN) operations. For the deep learning methods, we select the approaches designed to operate on text. These include gated recurrent unit (GRU), bidirectional gated recurrent unit (BiGRU), LSTM, BiLSTM, and convolutional neural network (CNN) [18][19].

To execute this experiment, we use the Google Translate API to translate all hacker forum post content into English for input into the selected benchmark algorithms. Performances were measured and evaluated against the proposed set of benchmarks in both settings. Both the Russian and French datasets remained untranslated for A-CLKT.

3) Experiment 2: A-CLKT vs Monolingual Models

Experiment 2 examines the performance of benchmark algorithms when the input text is not translated. Theoretically, this should retain the original semantics of the languages and enhance overall performance. To execute this experiment, we use the untranslated text as input for the same set of classical machine learning and deep learning algorithms. Like Experiment 1, we examined A-CLKT performances with English as the high-resource source language and either French or Russian as the low-resource target language.

4) Experiment 3: A-CLKT vs CLKT Alternatives

Experiment 3 compares the proposed A-CLKT against prevailing deep learning-based CLKT approaches. The first relies on a fully multi-lingual (FML) learning strategy. This approach trains a two-layer deep learning architecture without differentiating between languages. The second CLKT category employs a multi-task learning (MTL) strategy on multiple deep learning architectures with shared layers. Variations of this strategy included MTL-BiLSTM, MTL-LSTM, MTL-GRU, and MTL-BiGRU. Like Experiments 1 and 2, we examined A-CLKT performances with English as the high-resource source language and either French or Russian as low-resource language.

5) Performance Metrics

The imbalanced nature of our gold-standard datasets requires a careful training strategy with well-established performance metrics. Therefore, each algorithm is trained and tested using 5-fold cross-validation. We evaluate algorithm performances using accuracy, precision, recall, and F_1-Score [20]. Each uses a count of true positive (TP), false positive (FP), true negative (TN), and false negative (FN). In this context, TP is the quantity of correctly classified cyber threats, TN is the number of correctly classified non-cyber threats, FN is the number of cyber threats incorrectly classified as non-cyber threats, and FP is the number of non-cyber threats incorrectly classified as cyber threats. Each metric is computed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \qquad Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN}, \qquad F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}.$$

We also measured performance using a receiver operating characteristics (ROC) curve. ROC plots the true positive rate (y-axis) versus the false positive rate (x-axis). The area under the ROC curve determines the AUC score, which is a scalar metric ranging from 0.5 (random guess) and 1.0 (perfect performance). AUC quantifies the trade-offs between type I and type II errors. It is often a preferred metric when measuring the performance of algorithms operating on class-imbalanced datasets [21].

For each metric, we performed paired *t*-tests to evaluate statistical significance. Results were considered statistically significant for p-value thresholds of $p<0.001$, $p<0.01$, and $p<0.05$. Algorithms were implemented in Python with the Keras and Scikit-Learn packages on a single Ubuntu workstation with an Intel 3.30 GHz CPU, and a GeForce GTX Graphical Processing Unit (GPU) with 1,280 Cuda cores and six GB of GPU memory. Full model specifications are available in our publicly accessible GitHub repository at https://github.com/mohammadrezaebrahimi/A-CLKT .

V. Results and Discussion

A. Experiment 1 Results: A-CLKT vs Machine Translation-based Approaches

Table III summarizes A-CLKT's performance for each language against benchmark methods that rely on machine translation as input. The top-performing algorithm is highlighted in boldface.

TABLE III. Summary of A-CLKT Performance Against Machine Translation-based Models (Note: *** p<0.001, ** p<0.01, * p<0.05)

Russian Dataset						
Category	**Method**	**Acc.**	**Prec.**	**Recall**	**F_1**	**AUC**
Classical Machine Learning	k-NN	0.5549 ***	0.607	0.2857 ***	0.3526 **	0.5419 ***
	SVM	0.6437 ***	0.648	0.4460 ***	0.5245 ***	0.6232 ***
	NB	0.6282 **	0.5934 *	0.66	0.6231 *	0.6307 ***
	RF	0.6465 **	**0.6993**	0.4426 **	0.5362 **	0.6379 ***
Deep Learning Methods	LSTM	0.6100 ***	0.5707 ***	0.7461	0.6448 **	0.6646 ***
	BiLSTM	0.5920 ***	0.5297 ***	0.7277	0.6117 ***	0.6444 ***
	GRU	0.5908 ***	0.5842 *	**0.7592**	0.6314 ***	0.6470 ***
	BiGRU	0.6195 ***	0.5933 ***	0.6868	0.6112 ***	0.6270 ***
	CNN	0.6291 **	0.5887	0.5738	0.5703 *	0.6387 **
Proposed A-CLKT		**0.7562**	0.6832	0.7125	**0.6910**	**0.7999**
French Dataset						
Category	**Method**	**Acc.**	**Prec.**	**Recall**	**F_1**	**AUC**
Classical Machine Learning	k-NN	0.5491 ***	**0.8072**	0.2532 **	0.2870 ***	0.5414 ***
	SVM	0.6228 ***	0.6461	0.4988 ***	0.5585 ***	0.6198 ***
	NB	0.6193 ***	0.5826 **	0.7586	0.6557 *	0.6264 ***
	RF	0.6702 **	0.6973	0.5471 *	0.6082 *	0.6667 **
Deep Learning Methods	LSTM	0.6158 ***	0.5810 ***	0.7283	0.6429 **	0.6601 ***
	BiLSTM	0.6088 ***	0.5829 ***	0.7024	0.6314 **	0.6373 ***
	GRU	0.6386 ***	0.6225 *	0.7818	0.6907 *	0.6428 ***
	BiGRU	0.5842 ***	0.5577 ***	0.7077	0.6166 **	0.6162 ***
	CNN	0.6316 ***	0.6073 **	0.7144	0.6445 *	0.6811 ***
Proposed A-CLKT		**0.7452**	0.6836	**0.7634**	**0.7032**	**0.7720**

The table indicates that the proposed A-CLKT approach outperformed classical machine learning and deep learning-based methods operating on translated hacker forum text. Performances were statistically significant on the most comprehensive metrics of F_1-score and AUC. The consistency of these results for both the Russian and French datasets indicates that translating hacker forum text to English loses semantics within the language. Ultimately, this loss of information affects the ability of the output classifier to delineate between cyber threats and benign postings within hacker forums.

B. Experiment 2 Results: A-CLKT vs Monolingual Models

Table IV summarizes A-CLKT's performance for each language against methods that use untranslated text as input. The top-performing algorithm is highlighted in boldface.

TABLE IV. SUMMARY OF A-CLKT PERFORMANCE AGAINST MONOLINGUAL MODELS (NOTE: *** p<0.001, ** p<0.01, * p<0.05)

Russian Dataset						
Category	**Method**	**Acc.**	**Prec.**	**Recall**	**F_1**	**AUC**
Classical Machine Learning	k-NN	0.7627	0.2000 *	0.0105 ***	0.0200 ***	0.5032 ***
	SVM	0.78	0.5745	0.2759 ***	0.3616 ***	0.6083 ***
	NB	0.6229 ***	0.2981 ***	0.4487 ***	0.3542 ***	0.5632 ***
	RF	**0.7828**	**0.75**	0.1258 ***	0.2141 ***	0.5571 ***
Deep Learning Methods	LSTM	0.8000	0.6791	0.6028 **	0.6324 *	0.7627 *
	BiLSTM	0.7492 *	0.5850 *	0.6691	0.6169 **	0.7354 *
	GRU	0.6883 *	0.5444	0.7289	0.6056 *	0.7128 *
	CNN	0.7029	0.5153 **	0.6433 *	0.5574 ***	0.7321 *
	BiGRU	0.7308 *	0.5343 ***	0.5695 **	0.5492 ***	0.6940 **
Proposed A-CLKT		0.7562	0.6832	**0.7125**	**0.6910**	**0.7999**
French Dataset						
Category	**Method**	**Acc.**	**Prec.**	**Recall**	**F_1**	**AUC**
Classical Machine Learning	k-NN	0.7882	0.4	0.0486 ***	0.0864 ***	0.5208 ***
	SVM	**0.8117**	0.7	0.1655 ***	0.2607 **	0.5790 ***
	NB	0.6647 *	0.3211 ***	0.4864 *	0.3768 ***	0.6038 **
	RF	0.7999	**0.7**	0.1222 ***	0.2016 ***	0.5576 ***
Deep Learning Methods	LSTM	0.8000	0.6791	0.6028 **	0.6324 *	0.7627 *
	BiLSTM	0.7492 *	0.5850 *	0.6691	0.6169 **	0.7354 *
	GRU	0.6883 *	0.5444	0.7289	0.6056 *	0.7128 *
	CNN	0.7029	0.5153 **	0.6433 *	0.5574 ***	0.7321 *
	BiGRU	0.7308 *	0.5343 ***	0.5695 **	0.5492 ***	0.6940 **
Proposed A-CLKT		0.7452	0.6836	**0.7634**	**0.7032**	**0.7720**

Similar to Experiment 1, the A-CLKT outperformed benchmark methods operating on untranslated hacker forum text for both the Russian and French datasets. These differences were statistically significant for recall, F_1, and AUC. These results indicate that leveraging the untranslated knowledge across languages helps improve overall cyber threat detection.

C. Experiment 3 Results: A-CLKT vs. CLKT Alternatives

Table V summarizes A-CLKT's performance for each language against prevailing CLKT methods. The top-performing algorithm is highlighted in boldface.

TABLE V. SUMMARY OF A-CLKT PERFORMANCE AGAINST CLKT ALTERNATIVES (NOTE: *** p<0.001, ** p<0.01, * p<0.05)

Dataset	**Method**	**Acc.**	**Prec.**	**Recall**	**F_1**	**AUC**
Russian	FML-CNN	0.6611 *	0.4995 **	**0.7650**	0.5897 **	0.7018 **
	MTL-LSTM	0.6981	0.5493 *	0.6782	0.6030	0.6924 **
	MTL-GRU	0.5985 *	0.4465 ***	0.6898	0.5273* **	0.6355 ***
	MTL-BiLSTM	0.6868	0.5541	0.6091	0.5477*	0.6834 *
	MTL-BiGRU	0.6453 **	0.4536 **	0.6391	0.5205* **	0.6403 ***
	Proposed A-CLKT	**0.7562**	**0.6832**	0.7125	**0.6910**	**0.7999**
French	FML-CNN	**0.7520**	**0.7143**	0.5429 **	0.5682* **	0.6127 ***
	MTL-LSTM	0.6231	0.4753 *	0.7250	0.5333* *	0.6532 ***
	MTL-GRU	0.7231	0.5400 *	0.7195	0.6050* *	0.6886 **
	MTL-BiLSTM	0.7231	0.5345 **	0.6333	0.5518* *	0.6476 **
	MTL-BiGRU	0.7077	0.5165 **	0.7444	0.5954* *	0.6884 **
	Proposed A-CLKT	0.7452	0.6836	**0.7634**	**0.7032**	**0.7720**

Experiment 3 results suggest that A-CLKT's adversarial learning approach creates a more robust and comprehensive representation of English and non-English hacker forum text than its FML and MTL counterparts in detecting cyber threat in both Russian and French hacker forums based on F_1-score and AUC. Similar to the first two experiments, these differences were statistically significant. This indicates that the adversarial learning procedure systematically removes features that are less relevant to creating a language invariant representation. In contrast, the benchmark approaches may include them, thus causing a decrease in overall performance.

VI. CONCLUSION AND FUTURE DIRECTIONS

Despite cybersecurity's importance, the quantity and severity of cyber-attacks are on an unfortunate uptick. Many cybersecurity professionals are closely examining the international Dark Web to proactively pinpoint potential cyber threats. Despite its potential, the Dark Web contains hundreds of thousands of non-English posts. This limits an analyst's ability to pinpoint cyber threats in a scalable and automated fashion.

In this work, we aimed to take an important step in advancing multi-lingual cyber threat detection capabilities. Specifically, we designed a novel A-CLKT approach. A-CLKT formulates an innovative adversarial learning procedure to automatically learn language invariant representations across two languages. A series of benchmark experiments illustrated how A-CLKT outperformed classical machine learning, deep learning, and CLKT algorithms in detecting cyber threats in Russian and French hacker forums.

There are several promising directions for future research. First, while we formulated our task as binary classification, future studies can aim group the posts into finer-grained output labels (e.g., type of hacking tool). Second, future studies can explore if including additional generators to represent multiple languages improves overall performance. Finally, future work can explore how the proposed approach operates on other Dark Web platforms. Each direction can help cyber-analysts proactively identify cyber-threats in the international Dark Web.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation (NSF) under Grants SES-1314631 (SaTC SBE), ACI-1443019 (DIBBs), CNS-1936370 (SaTC CORE), and CNS-1850362 (CRII SaTC).

REFERENCES

[1] H. Chen, *Dark web: Exploring and data mining the dark side of the web*. New York: Springer, 2012.

[2] P.-Y. Du *et al.*, "Identifying, Collecting, and Presenting Hacker Community Data: Forums, IRC, Carding Shops, and DNMs," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*, Miami, FL, 2018.

[3] E. Nunes *et al.*, "Darknet and deepnet mining for proactive cybersecurity threat intelligence," in *IEEE Conference on Intelligence and Security Informatics (ISI)*, Tucson, AZ, 2016, pp. 7–12.

[4] N. Arnold *et al.*, "Dark-Net Ecosystem Cyber-Threat Intelligence (CTI) Tool," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2019, pp. 92–97, doi: 10.1109/ISI.2019.8823501.

[5] W. Li, H. Chen, and J. F. Nunamaker Jr, "Identifying and Profiling Key Sellers in Cyber Carding Community: AZSecure Text Mining System," *Journal of Management Information Systems*, vol. 33, no. 4, pp. 1059–1086, 2016, doi: 10.1080/07421222.2016.1267528.

[6] N. Tavabi, P. Goyal, M. Almukaynizi, P. Shakarian, and K. Lerman, "Darkembed: Exploit prediction with neural language models," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[7] M. Schäfer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti, and V. Lenders, "BlackWidow: Monitoring the Dark Web for Cyber Security Information," in *International Conference on Cyber Conflict (CyCon)*, 2019, vol. 900, pp. 1–21.

[8] J. Grisham, S. Samtani, M. Patton, and H. Chen, "Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*, Beijing, China, 2017, pp. 13–18.

[9] W. Li and H. Chen, "Identifying top sellers in underground economy using deep learning-based sentiment analysis," in *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*, 2014, pp. 64–67.

[10] S. Samtani, R. Chinn, H. Chen, and J. F. Nunamaker Jr, "Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence," *Journal of Management Information Systems*, vol. 34, no. 4, pp. 1023–1053, 2017.

[11] S. Samtani, R. Chinn, and H. Chen, "Exploring hacker assets in underground forums," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*, Baltimore, MD, 2015, pp. 31–36.

[12] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016, doi: 10.1186/s40537-016-0043-6.

[13] M. Abdalla and G. Hirst, "Cross-Lingual Sentiment Analysis Without (Good) Translation," in *The 8th International Joint Conference on Natural Language Processing (IJCNLP)*, Taiwan, 2017, pp. 506–515.

[14] N. Li, S. Zhai, Z. Zhang, and B. Liu, "Structural Correspondence Learning for Cross-Lingual Sentiment Classification with One-to-Many Mappings," in *AAAI Conference on Artificial Intelligence*, San Francisco, 2017, pp. 3490–3496.

[15] M. Wang and W. Deng, "Deep Visual Domain Adaptation: A Survey," *Neurocomputing*, 2018.

[16] I. Goodfellow *et al.*, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[17] R. Johnson and T. Zhang, "Supervised and Semi-supervised Text Categorization Using LSTM for Region Embeddings," in *International Conference on Machine Learning (ICML)*, New York, NY, 2016, vol. 48, pp. 526–534.

[18] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT Press Cambridge, 2016.

[19] Y. Goldberg, "Neural Network Methods for Natural Language Processing," *Synthesis Lectures on Human Language Technologies*, vol. 10, no. 1, pp. 1–309, 2017.

[20] M. Ebrahimi, M. Surdeanu, S. Samtani, and H. Chen, "Detecting Cyber Threats in Non-English Dark Net Markets: A Cross-Lingual Transfer Learning Approach," in *IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2018, pp. 85–90.

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer series in statistics New York, 2017.