Detecting Cyber Threats in Non-English Dark Net Markets: A Cross-Lingual Transfer Learning Approach

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

Mihai Surdeanu
Department of Computer Science
University of Arizona
Tucson, Arizona, USA
msurdeanu@email.arizona.edu

Sagar Samtani
Department of Information Systems and Decision Sciences
University of South Florida
Tampa, Florida, USA
ssamtani@usf.edu

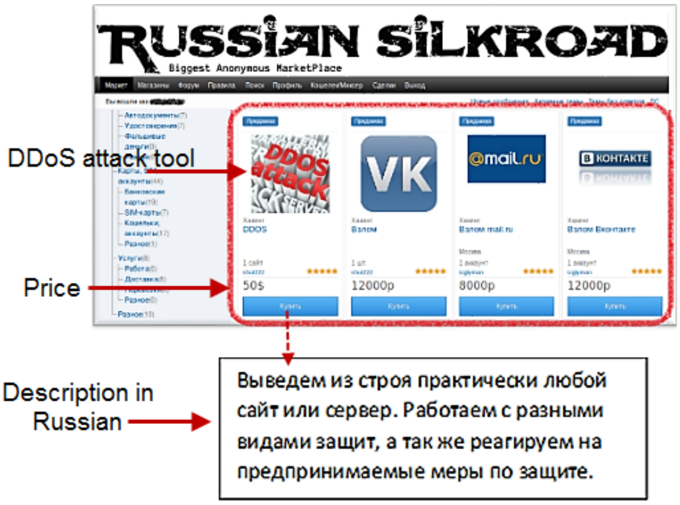
Hsinchun Chen
Department of Management Information Systems
University of Arizona
Tucson, Arizona, USA
hchen@eller.arizona.edu

***Abstract*—Recent advances in proactive cyber threat intelligence rely on early detection of cyber threats in hacker communities. Dark Net Markets (DNMs) are growing platforms in hacker community that provide hackers with highly-specialized tools and products which may not be found in other platforms. While text classification techniques have been used for cyber threat detection in English DNMs, the task is hindered in non-English platforms due to the language barrier and lack of ground-truth data. Current approaches use monolingual models on machine translated data to overcome these challenges. However, the translation errors can deteriorate the classification results. The abundance of data in English DNMs can be leveraged in learning non-English threats without using machine translation. In this study, we show that a deep cross-lingual model that can jointly learn the common language representation from two languages, significantly outperforms a monolingual model learned on machine translated data for identifying cyber threats in non-English DNMs. Unlike most studies, our approach does not require any external data source such as bilingual word embeddings or bilingual lexicons. Our experiments on Russian DNMs show that this approach can achieve better performance than state-of-the-art methods for non-English cyber threat detection in malicious hacker community.**

***Keywords*—Dark Net Markets, cyber threat, deep learning, cross-lingual transfer learning**

I. Introduction

Proactive Cyber Threat Intelligence (CTI) aims to mitigate the risk of cyber attacks by detecting emerging cyber threats in the hacker community [1]. Dark Net Markets (DNMs), hacker forums, carding shops, and Internet-Relay-Chat (IRC) comprise the vast online hacker community. Among them, Dark Net Marketplaces (DNMs) are an integral and unique part of this broad ecosystem in the sense that their anonymity and profitability provide an environment conducive to cybercriminal activities. DNMs host purchasable highly-specialized products (listings) that are not available in other platforms (e.g., ransomware, keyloggers, SQL Injection tools, DDoS attack tools, stolen account information, and hacked personal credentials). These malicious products are viewed as threats to cybersecurity since they are often used by hackers to conduct cyber attacks. Since 2013, the number of language-specific DNMs have increased [2]. While English is the dominant language, Russian, French, and Italian are common non-English DNMs [3]. Based on the DNMs listed in deepdotweb.com, a popular website providing up-to-date directory of DNMs [4], almost 56% of platforms are English, 19% are Russian, 12.5% are French, and the rest are Italian. Non-English DNMs reflect different geopolitical regions and vary in malicious products. In particular, while English DNMs are geared towards general hacking contents, the Russian DNMs offer specialized hacking services such as personalized email hacking, call flooding, and Distributed Denial of Service (DDoS) attacks (Fig. 1). Therefore, detecting cyber threats in non-English DNMs is crucial to providing global insight across hacker community.



Выведем из строя практически любой сайт или сервер. Работаем с разными видами защит, а так же реагируем на предпринимаемые меры по защите.

Fig. 1. Listings of Hacking Tools in a Russian DNM.

Despite their CTI value, DNMs also contain many non-cyber threat products (e.g., digital goods and drugs). Manual cyber threat detection is impractical due to increasing number of new products. Researchers have adopted text classification to automate threat detection [5][6]. Text classification models require labeled data for training. While labeled data in English is often available, the language barrier results in limited labeled data in non-English DNMs which hinders cyber threat detection. Current studies use machine translation (MT) to tackle this challenge [7]–[9]. However, informal, non-grammatical and hacker-specific language causes translation

978-1-5386-7848-0/18/$31.00 ©2018 IEEE

errors. Machine translation errors often propagate to the model and deteriorate threat detection performance [10][11]. Deterioration of threat detection performance results in high false negative rate which leads to overlooking potentially important threats (e.g., missing a DDoS attack tool), and high false positive rate which leads to suggesting non-threats (e.g., a book about hacking) as threats. These issues motivate developing models that transfer knowledge from high-resource languages (e.g., English) to low-resource ones (e.g., Russian) without relying on MT. Transfer learning is a field of machine learning which is concerned with transferring the learned knowledge from a source domain or task to a target domain or task [12]. Here we consider an application of transfer learning known as Cross-Lingual Knowledge Transfer (CLKT), which aims to transfer the knowledge between two languages. Deep learning has gained success as a method to automatically extract salient features that are transferrable among languages and also provides an efficient way to transfer the knowledge via sharing network layers and parameters [13]. Among deep learning architectures, Bidirectional Long Short-Term Memories (BiLSTMs) have the capability of using word orders [14] and capturing temporal patterns. This makes them suitable for capturing the common language representation from different languages [15].

In this study, we propose a novel supervised knowledge transfer method for detecting cyber threats in non-English DNMs. This method leverages the labeled data in English DNMs in conjunction with the limited labeled data in target non-English DNMs simultaneously and learns a shared BiLSTM to capture common hacker language representation, which helps transferring the knowledge learned from English to non-English DNMs. Our approach differs from other deep CLKT approaches in the sense that it does not need any external resources such as mono or bilingual word-embeddings, neither machine translation. Also most prior non-cybersecurity studies assume there is no labeled data in the target language and therefore address unsupervised knowledge transfer. Given that in cybersecurity applications some labeled data is often available in the target language, using supervised approaches can lead to better performance as opposed to unsupervised methods. Our method reduces false positive and false negative rates without relying on MT and significantly outperforms the state-of-the-art methods used for detecting non-English cyber threats in CTI.

II. Literature Review

Given our research objective, four areas of research are reviewed. First, we investigate cyber threat detection in hacker community to examine the past efforts on detecting cyber threats in English and non-English platforms. Second, we review deep transfer learning and third, we examine its application, deep CLKT, to inform the development of a knowledge transfer method for cyber threat detection in non-English DNMs. Finally, we describe BiLSTMs as a promising architecture for extracting and sharing the common language representation.

A. Cyber Threat Detection in Hacker Community

Hackers share knowledge and malicious tools through DNMs, hacker forums, carding shops, and IRC channels [16]. We summarize the prior work on these four platforms based on the supported languages and the approach used to deal with non-English platforms in order to gain a comprehensive insight into the state of the art in cyber threat detection in hacker community. Prior work on threat detection in hacker communities falls into two main categories: (1) monolingual approach and (2) machine translation-based approach. The former aims to build separate models in each language with language-specific features, while the latter uses machine translation to translate the non-English data to English and learns a monolingual model on the translated data subsequently.

The majority of work in the first category is centered on English platforms. An explorative analysis of data sources in the dark web is conducted in [16] via monolingual keyword search and information retrieval on five hacker forums, eight IRC channels, and four carding shops in English and Russian. In another work [17], authors use monolingual recurrent neural networks to discover hacker jargons and terms in English forums. Similarly in [5], authors employ semi-supervised labeling and monolingual SVM classifiers to tackle threat detection in ten English DNMs. In [6], malicious product grouping in 17 English DNMs was presented via K-Means clustering.

In the second category more work has been done on non-English platforms. In [7], Support Vector Machine (SVM) was used in a monolingual setting to classify malware source codes on eight forums in English and Russian. In their approach, Russian data was machine translated to English using Google Translate. Similarly, in [8] authors used Maximum Entropy classification and Recursive Neural Networks to detect and rate top sellers in eight carding shops in English and Russian using Google Translate. Their approach uses SentiTreeBank, an external pre-trained word embedding trained on customer reviews. Another machine translation-based approach is used in [9], which applies LSTM to detect mobile malware on four forums in English, Arabic, and Russian. Given that the methods in this category represent the state of the art in threat detection, their performance serve as baseline for our proposed method.

There are two key limitations associated with current cyber threat detection models in non-English platforms. First, training separate monolingual models on low-resource non-English languages is not practical [18][19] and cannot provide a global insight across DNMs with different languages. Second, erroneous machine translated technical text can affect system performance [10][11]. As a result, monolingual models applied on machine translated data may suffer from poor threat detection performance. Moreover, prior DNM studies only focus on English DNMs.

In light of these limitations, we review deep transfer learning and its extension in cross-lingual knowledge transfer to inform the development of a method that can transfer the learned knowledge from a high-resource language (i.e., English) to low-resource non-English DNMs.

B. *Deep Transfer Learning*

Transfer learning aims to leverage the knowledge obtained from a resource-rich task to solve a resource-deprived one in which there is not sufficient training data [12]. Recent progress in deep learning has revealed that layers and parameters in deep architectures can capture the underlying domain-invariant data representation [20]–[22]. Also, deep learning has shown promising results in automatically extracting important transferable features among tasks [13][23][24].

Achieving a domain-invariant representation and learning transferable features form the main idea of deep transfer learning and are mostly implemented via sharing layers and parameters among different tasks [13][25]. That is, sharing hidden layers in deep architectures facilitates using the knowledge from source domain to improve the performance in target domain [25].

C. *Deep Cross-Lingual Knowledge Transfer (CLKT)*

CLKT is intended to improve the learning in target task in low-resource language domain $D^{L'}$ by using the knowledge from a source task in high-resource language domain D^{L}. D^{L} and $D^{L'}$ encompass the feature space in languages L and L', respectively. The task can be any machine learning problem of interest. For instance, in the case of cross-lingual threat detection in DNMs, source and target tasks both are formulated as assigning label $y \in \{0,1\}$ to a product description from source or target language L (English) and L' (Russian). In this section, we review selected prior studies on deep CLKT for low-resource text based on the application context and the method used in each study.

CLKT has been successfully used in cross-lingual sentiment classification [26]–[28]. Different deep learning architectures including BiLSTM [26], adversarial networks [27], and stacked denoised autoencoders [28] have been used for capturing common language representation in sentiment classification. This area of research and cross-lingual threat detection are similar in the sense that both problems reduce to binary text classification. Cross-lingual knowledge transfer has also been used in fundamental NLP tasks such as Semantic Textual Similarity (STS) and language modeling. In [29], adversarial learning in conjunction with shared BiLSTMs was used to learn the bilingual sentence representation in identifying textual similarity. Also in [30], an LSTM language model with cross-lingual word embeddings was developed to improve the intrinsic quality of the language model.

Methods in [26]–[28] assume there is no labeled data in the target language (unsupervised knowledge transfer). However, since limited labels in target language are often available in cybersecurity applications, we design our method such that it utilizes the limited labels in target DNMs (supervised knowledge transfer). Also, the approach used in [29] requires monolingual word embeddings in source and target language as an auxiliary resource, unlike our method which does not require any external data sources.

D. *Bidirectional Long Short-Term Memory (BiLSTM)*

As noted, BiLSTMs have shown to be successful in capturing common language representations due to their ability in capturing word orders [14]. BiLSTM consists of two LSTM layers, each reading the input sequence in opposite directions (Fig. 2).

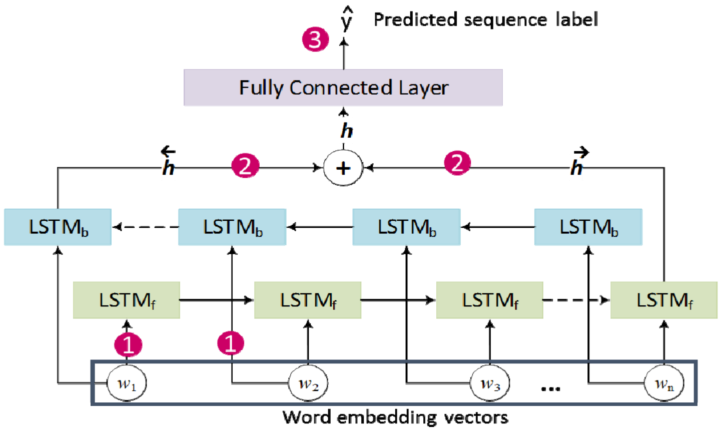


Fig. 2. Graphical Representation of Basic BiLSTM for Text Classification

In Fig. 2, the goal is to predict label $\hat{y}$ for the entire input sequence $\langle w_1, w_2, ..., w_n \rangle$ in which w_i denotes the i^{th} word embedding vector. In the first step, forward and backward LSTMs (subscripted by f and b) read word embeddings in parallel and generate hidden states $\overrightarrow{h} \in R^{d_h}$ and $\overleftarrow{h} \in R^{d_h}$ in which R^{d_h} represents a real-valued vector with size d_h. In the second step, the final hidden state $h \in R^{2 \times d_h}$ is obtained from concatenating the final hidden states $\overrightarrow{h}$ and $\overleftarrow{h}$ in forward and backward layers. The loss $(y - \hat{y})$ is calculated in step 3. Finally, steps 1-3 are repeated until the loss function is minimized.

E. *Research Gaps and Questions*

Several research gaps are identified from the literature review. Most DNM studies only identify threats in English DNMs. Hence, threats in non-English DNMs (e.g., Russian) are understudied, yet critically needed. Prior studies addressing multiple languages either use independent monolingual models or train monolingual models on machine translated data which can lead to poor classification performance on low-resource non-English DNMs. The following questions are posed to address the identified gaps:

- How can CLKT be leveraged for cyber threat detection in non-English DNMs without machine translation?
- How can the threat knowledge learned from English DNMs be transferred to non-English DNMs?

Motivated by these questions, we propose a novel transfer learning framework to conduct cross-lingual cyber threat detection in non-English DNMs. To our knowledge, this is the first framework that addresses the task in non-English DNMs.

III. Research Design

Our transfer learning-based cross-lingual cyber threat detection framework has three major components: data collection and preparation, bilingual testbed generation, and cross-lingual cyber threat detection and evaluation (Fig. 3).

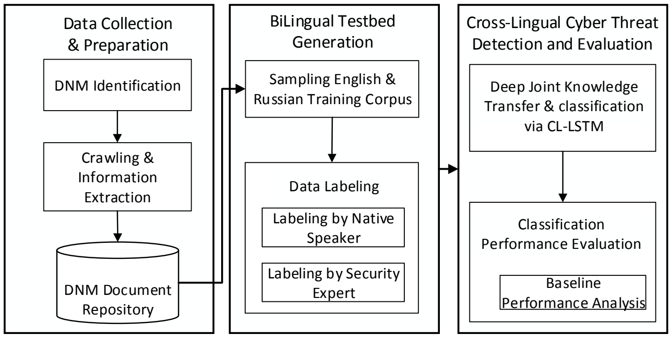

Data Collection & Preparation
DNM Identification
Crawling & Information Extraction
DNM Document Repository
BiLingual Testbed Generation
Sampling English & Russian Training Corpus
Data Labeling
Labeling by Native Speaker
Labeling by Security Expert
Cross-Lingual Cyber Threat Detection and Evaluation
Deep Joint Knowledge Transfer & classification via CL-LSTM
Classification Performance Evaluation
Baseline Performance Analysis

Fig. 3. Proposed Framework for Transfer Learning-Based Cross-Lingual Cyber Threat Detection

A. *Data Collection and Preparation*

Seven English DNMs and one Russian DNM were identified based on deepdotweb.com. A web spider employing techniques to avoid anti-crawling measures traversed each DNM to extract all product descriptions. 95,095 product descriptions were collected and parsed in a relational database. The product descriptions include cyber-threat related products (e.g., DDoS, ransomware, and keyloggers) and non-cyber threats (e.g., drugs, digital goods, books, and weapons).

B. *Bilingual Testbed Generation*

To facilitate the training of our proposed model, we randomly sampled the product descriptions in each language by preserving the ratio of cyber to non-cyber products and obtained 2,373 product listings including 1,821 English and 552 Russian products (Table I). English product description were manually labeled as cyber threats or non-threats by two cybersecurity experts. Russian products were manually labeled by a cybersecurity expert and a Russian speaker.

TABLE I. Summary of DNM Data Collection

DNM	# of listings	Language	# of Labeled Products	Year
Dream Market	39,473	English	1,821	2016, 2017
AlphaBay	25,118			
Hansa	14,149			
Silkroad3	1,683			
Minerva	683			
Apple Market	877			
Valhalla	12,192			
Russian Silkroad	920	Russian	552	2016
Total:	95,095	-	2,373	2016-2017

C. *Cross-Lingual Cyber Threat Detection*

As noted in the deep CLKT review, joint learning of shared hidden layers has shown to produce high-quality language representations. The resulting BiLSTM (without the fully connected layer) can be shared between two networks each training on a different language to capture the commonalities from both languages in the form of hidden states [15][29]. Motivated by this finding, we propose CL-LSTM (Cross-Lingual LSTM) which employs BiLSTM to jointly learn the common hacker language representation from English and non-English DNMs. Our method is inspired by a monolingual architecture in [31], which includes joint learning of a common representation from two classification tasks in English (i.e., sentiment classification and subjectivity analysis) using shared BiLSTMs in a multi-task manner with homogenous feature spaces. Unlike this study, our approach deals with heterogeneous feature spaces in two different languages and is not concerned with learning from different tasks. Furthermore, [31] uses Word2Vec English word embeddings as external resources, while the word embeddings in our method are randomly initialized and are learned during the training process.

We developed CL-LSTM for Russian cyber threat detection. The architecture and key operations of CL-LSTM are shown in Fig. 4. There are three layers in this architecture: a language-independent shared BiLSTM layer to jointly capture the common representation from English and Russian, and two language-specific LSTM layers that interpret the common representation for threat detection in Russian and English separately (shown by LSTM^{ru} and LSTM^{en}). While English and Russian labeled data (i.e., supervision signals) are different for each language-specific layer the weights and structure of shared BiLSTM is shared and therefore is the same for both languages.

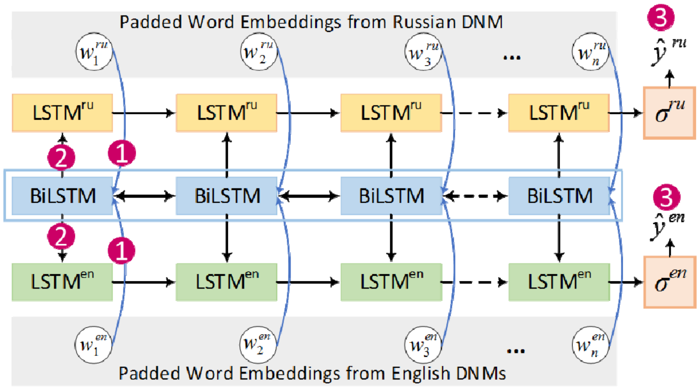


Fig. 4. Graphical Illustration of CL-LSTM for Joint Cross-Lingual Knowledge Transfer from English to Russian.

As shown in Fig. 4, in step 1 (represented in red circle), the shared BiLSTM layer reads word embeddings of products in English and Russian DNMs in parallel. In step 2, the hidden state vectors emitted by the shared layer at each time-step are fed to language-specific layers. In step 3, the class labels for the products in each language are predicted independently via softmax functions σ^{en} and σ^{ru}. Lastly, the loss values are calculated for English and Russian separately and gradient errors propagate to the shared layer. These steps repeat until the loss is minimized or a stop condition (e.g., maximum number of training epochs) is met.

The shared bidirectional layer (BiLSTM cells in Fig. 4) is composed of a forward and a backward LSTM as already shown in Fig. 2. The only difference with the conventional

BiLSTM is that the weight matrices are shared between Russian- and English-specific layers. While different specifications are available for BiLSTMs we implemented the specification in [32] (Eq. 1-4).

$$h_t = o * \tanh(C_t) \tag{1}$$

$$C_t = C_{t-1} * f_t + i_t * z_t \tag{2}$$

$$g_t = \sigma\left(W_{xg}^{shared} x_t + U_{hg}^{shared} h_{t-1}\right) + b_i^{shared} \tag{3}$$

$$z_t = \tanh\left(W_{xz}^{shared} x_t + U_{hz}^{shared} h_{t-1}\right) + b_z^{shared} \tag{4}$$

h_t and C_t are hidden state and cell state at time *t*. Also *g* represents either of input (*i*), forget (*f*), and output (*o*) gate vectors and * denotes component-wise vector multiplication. W_{xg}^{shared} denotes shared weight matrices from input vector x_t to input, forget or output gates. Similarly, U represents weight matrices between hidden state vectors and the gates. b^{shared} denotes the common bias terms. z_t is the potential update computed as in simple recurrent neural network. σ can be any non-linearity. The same specification applies to the backward LSTM. The final output from each cell in the shared layer is generated by concatenating hidden states in forward and backward LSTMs [32].

$$y_t^{shared} = \vec{h}_t \oplus \overleftarrow{h}_t \tag{5}$$

The learning procedure of our CL-LSTM is summarized in Algorithm 1. Cross-entropy loss [29] was used as loss function and was minimized by Adam optimizer [33]. Also *tanh* and *sigmoid* were used as activations.

Algorithm 1. CL-LSTM Learning Procedure for Russian DNMs.

```
Inputs: Word embedding sequences $W^{en}$ and $W^{ru}$ in which
$W_i^{en} = \left\langle w_1^{en}, w_2^{en}, \dots, w_n^{en} \right\rangle$, $W_i^{ru} = \left\langle w_1^{ru}, w_2^{ru}, \dots, w_n^{ru} \right\rangle$ represent
the sequence for $i^{th}$ product description in the domain of source and target
DNM ( $D^{en}$, $D^{ru}$ ), respectively.
Output: Predicted class label for products from target DNM ( $\hat{y}^{ru}$ ).
while $CrossEntropy(y, \hat{y}^{ru})$ is minimized or stop condition is met, do
 for each batch of word embedding sequence $W_i^{en}$ and $W_i^{ru}$ do
    - Compute the hidden states $h_t$ for shared biLSTM by Eq. 1-4.
    - Feed $h_t$ as input to language-specific LSTMs and generate the final
      hidden state from the last cell in each layer to obtain $\hat{y}^{ru}$.
    - Calculate the error gradients via $CrossEntropy(y, \hat{y}^{ru})$.
    - Update shared layer weights ( $W^{shared}$ and $U^{shared}$ ) and language-
      specific layers' weights through propagating the error gradients.
 end for
end while
return $\hat{y}^{ru}$
```

D. Performance Evaluation

As noted in the literature review section, machine-translation based methods represent the state of the art in cyber threat detection. Accordingly, to compare the results, we applied the approach used in [7][9] to our dataset. We denote these two methods as SVM + MT and LSTM + MT, respectively. The methods used in [29][30][8] leverage pre-trained word embeddings and therefore are excluded from our evaluation. Similarly, the unsupervised approaches mentioned in CLKT review [26]–[28] were excluded for a fair comparison.

The training and testing partitions were constructed via random assignment (80% to 20%). Hyperparameters (e.g., activation type, batch size) were tuned through 2-fold cross-validation. For deep learning models, we ran each test 10 times and averaged the results. BiLSTM layers in all models have equal number of cells (i.e., 100). Accuracy, precision, recall, and F1-score have already been used as cyber threat detection performance measures in the literature [7]–[9]. To further evaluate our method, we also compare the area under ROC curve (AUC). The statistical significance of the results was calculated by paired t-test [34].

IV. Results and Discussion

We compare our method to the benchmarks in terms of accuracy, precision, recall, and F1-score (Table II and Fig. 5)

TABLE II. Evaluating CL-LSTM Against Baselines

Method Description	Accuracy	Precision	Recall	F1-score
SVM + MT	95.15***[a]	96.15**	86.21***	90.91***
LSTM + MT	96.50*	97.88	89.66*	93.52**
BiLSTM + MT	94.17***	92.18**	87.24*	89.45***
CL-LSTM	**97.50**	**98.90**	**92.28**	**95.39**

[a] P-values significant at 0.05:*, 0.01:**, 0.001:***.

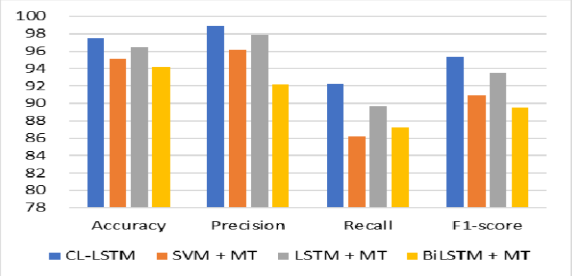


Fig. 5. CL-LSTM Performance Comparison

CL-LSTM outperforms all methods in accuracy, recall, and F1-score by statistically significant margins. Fig. 6 compares the AUC for deep learning benchmarks indicating that CL-LSTM can correctly detect more threats, while it reduces the number of false positives.

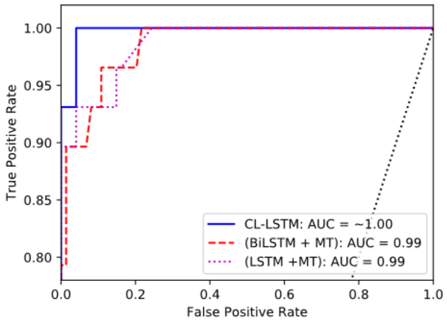


Fig. 6. Comparing the AUC for deep learning benchmarks

V. Conclusion and Future Directions

In this study, we proposed a novel transfer learning-based cyber threat detection framework for non-English DNMs using deep CLKT. We showed that threat knowledge learned from English DNMs can be transferred to Russian DNMs. Our approach jointly learns the common hacker-specific representation from Russian and English DNMs and outperforms baselines without relying on machine translation. Our framework advances proactive CTI by bridging the gap caused by language barrier in non-English DNMs and can help CTI professionals gain a better insight about cyber threats in foreign language DNMs. Future research is needed on developing methods to handle very short product descriptions at the character level. Validating the framework on other platforms (e.g., hacker forums) and other target languages (e.g., Arabic, Chinese) is another promising research direction.

Acknowledgment

This work was supported by the National Science Foundation (NSF) under Grant No. SES-1314631 (SaTC) and ACI-1443019 (DIBBs).

References

[1] J. Robertson *et al.*, *Darkweb Cyber Threat Intelligence Mining*. Cambridge University Press, 2017.

[2] J. Broséus, D. Rhumorbarbe, M. Morelato, L. Staehli, and Q. Rossy, "A geographical analysis of trafficking on a popular darknet market," *Forensic science international*, vol. 277, pp. 88–102, 2017.

[3] Europol and European Monitoring Centre for Drugs and Drug Addiction, "Drugs and the darknet: perspectives for enforcement, research and policy," 2017.

[4] "DeepDotWeb." [Online]. Available: http://www.deepdotweb.com. [Accessed: 08-Jun-2018].

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

[6] E. Marin, A. Diab, and P. Shakarian, "Product offerings in malicious hacker markets," in *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on*, 2016, pp. 187–189.

[7] 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.

[8] W. Li, H. Chen, and J. F. N. 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.

[9] 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 *Intelligence and Security Informatics (ISI), 2017 IEEE International Conference on*, 2017, pp. 13–18.

[10] S. Duek and S. Markovitch, "Automatic Generation of Language-Independent Features for Cross-Lingual Classification," *arXiv preprint arXiv:1802.04028*, 2018.

[11] S. C. AP *et al.*, "An autoencoder approach to learning bilingual word representations," in *Advances in Neural Information Processing Systems*, 2014, pp. 1853–1861.

[12] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, May 2016.

[13] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*, 2015, pp. 1225–1237.

[14] R. Johnson and T. Zhang, "Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings," in *ICML*, 2016, pp. 526–534.

[15] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[16] V. Benjamin, W. Li, T. Holt, and H. Chen, "Exploring threats and vulnerabilities in hacker web: Forums, IRC and carding shops," in *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*, 2015, pp. 85–90.

[17] V. Benjamin and H. Chen, "Developing understanding of hacker language through the use of lexical semantics," in *Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on*, 2015, pp. 79–84.

[18] L. Duong, H. Kanayama, T. Ma, S. Bird, and T. Cohn, "Learning Crosslingual Word Embeddings without Bilingual Corpora," *arXiv preprint arXiv:1606.09403*, 2016.

[19] X. Wan, "Co-training for cross-lingual sentiment classification," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*, 2009, pp. 235–243.

[20] M. Long, J. Wang, Y. Cao, J. Sun, and S. Y. Philip, "Deep learning of transferable representation for scalable domain adaptation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2027–2040, 2016.

[21] M. Chen, Z. Xu, and K. Q. Weinberger, "Marginalized Denoising Autoencoders for Domain Adaptation," in *Proceedings of the 29th International Conference on Machine Learning*, 2012, pp. 767–774.

[22] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 513–520.

[23] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

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

[25] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," in *Proceedings of EMNLP*, 2016.

[26] M. S. Rasooli, N. Farra, A. Radeva, T. Yu, and K. McKeown, "Cross-lingual sentiment transfer with limited resources," *Machine Translation*, pp. 1–23, 2017.

[27] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *arXiv preprint arXiv:1606.01614*, 2016.

[28] J. T. Zhou, S. J. Pan, I. W. Tsang, and Y. Yan, "Hybrid Heterogeneous Transfer Learning through Deep Learning.," in *AAAI*, 2014, pp. 2213–2220.

[29] J. Tian *et al.*, "An Adversarial Joint Learning Model for Low-Resource Language Semantic Textual Similarity," in *Advances in Information Retrieval*, Cham, 2018, pp. 89–101.

[30] O. Adams, A. Makarucha, G. Neubig, S. Bird, and T. Cohn, "Cross-lingual word embeddings for low-resource language modeling," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, vol. 1, pp. 937–947.

[31] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," in *International Joint Conferences on Artificial Intelligence*, New York city, 2016, pp. 2873–2879.

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

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, San Diego, CA, 2015.

[34] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.