BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2025
http://excel.fit.vutbr.cz

Transformers: The Detection of Malicious Domains

Filip Bučko

Abstract

Traditional detection methods for malicious domain names depend on **time-consuming feature engineering**, which allows attackers to evade detection. This paper utilizes **transformer neural networks** for **featureless detection** of **malware, phishing,** and **DGA** domains, learning directly from domain data. The manual creation of discriminative features is a significant bottleneck in security systems and often fails to generalize to novel attack patterns. Transformer networks offer a solution by automatically learning relevant features from sequential data, reducing this reliance on expert knowledge. A transformer model development process involved experimenting with various transformer architectures and tokenization strategies for **domain names, RDAP, DNS, and IP-derived geolocation** data, achieving **strong F1-scores**, with up to **98.6% for DGA** domains, **95% for malware**, and **98% for phishing**. The resulting featureless approach offers a resilient alternative to manual feature extraction, improving malicious domain detection.

xbucko05@vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Malicious domains are a key component of many cyberattacks (malware, phishing, DGA-based malware). Traditional detection via manual feature engineering is slow and easily evaded. This research presents a featureless transformer network approach to counter the increasing volume and complexity of cyberattacks. Learning directly from domain data offers robust, adaptable, and efficient detection across diverse datasets.

2. Related Works

The detection of malicious domains has been addressed through traditional techniques and modern machine learning methods. Conventional methods use similarity metrics such as Levenshtein distance, Jaccard index, and Kullback-Leibler divergence [1, 2, 3, 4].

Machine learning detects patterns using feature-based approaches with models such as random forests, decision trees, and gradient-boosted trees [5, 6, 7, 8, 9]. In contrast, deep learning extracts hierarchical representations directly from the input, utilizing architectures like convolutional (CNN) and long short-term memory (LSTM) networks for detection [10, 11, 12]. Natural Language Processing (NLP) treats domains as text, employing tokenization, syntactic analysis, and semantic analysis techniques to identify malicious patterns [13, 14].

Contributions Lightweight transformer models enable effective malicious domain identification from raw text and metadata, eliminating the reliance on manual feature engineering. A single, efficient architecture achieves state-of-the-art accuracy across diverse DNS-related data with real-time deployment potential, offering a more straightforward and adaptable foundation for DNS threat detection.

3. Solution

The training and evaluation of models utilized a benign domain dataset (830,344 entries from CESNET and Cisco Umbrella), phishing domains (164,425 obtained from Phishtank and OpenPhish, filtered with VirusTotal), a malware domain set (100,809 from ThreatFox, The FireBog, and MISP, with VirusTotal verification), and DGA domains (230,070 sourced from DGArchive). Due to the temporary nature and limited data of DGA domains, model training focused solely on domain names for their detection.

3.1 Data Selection and Preprocessing

Considering the capabilities of transformer architectures to process textual data effectively, four main data categories were selected.

- **Domain Names**
- **RDAP Records**
- **DNS Records**
- **Geographical Data**

Uninformative or redundant attributes, such as website addresses with mostly missing data, were discarded based on statistical tests. The kept attributes were then tokenized for transformer use with [CLS] and [SEP] markers.

3.2 Domain Name Analysis

The limited token sequence length of domain names makes them suitable for efficient initial architecture tuning experiments. For this purpose, the following architectures were utilized:

- **Pre-trained Transformers** – DistilBERT, BERT variants, ELECTRA, ALBERT-base, and MobileBERT, with sequence lengths adapted per model.
- **Custom Architecture** – Explored N-gram and character-level tokenization strategies.

DistilBERT achieved optimal performance, effectively balancing accuracy and computational efficiency. Consequently, this architecture was fine-tuned for malware, phishing, and DGA detection.

3.3 Extended Feature-Specific Models

RDAP Analysis: RDAP analysis focused on registrant, registrar, and admin contact info (email, whois_server, phone). Removing redundant flags improved model generalization. The RDAP model outperformed domain-only methods, showing the value of registrar data.

DNS Record Processing: Text-rich DNS records (MX, NS, SOA) were used after removing duplicates and low-value entries like repeated zone_SOA.

Geographical data: The transformer's input included country, region, city, and timezone, obtained by geolocating IP addresses from DNS records (A, AAAA, CNAME) via GeoLite2 databases.

4. Results

Tables 1 (domain names), 2 (RDAP), 3 (DNS), and 4 (Geographical information) summarize the models' performance across various data sources, with a focus on Accuracy (Acc), Precision (Prec), Recall (Rec), and the F1-score.

Task	Acc	Prec	Rec	F1
DGA	0.9855	0.9793	0.9921	0.9857
Malware	0.8945	0.8888	0.9003	0.8945
Phishing	0.9404	0.9515	0.9290	0.9401

Table 1. Domain Name Performance

Task	Acc	Prec	Rec	F1
Malware	0.9596	0.9544	0.9486	0.9515
Phishing	0.9802	0.9853	0.9811	0.9832

Table 2. RDAP Data Performance

Task	Acc	Prec	Rec	F1
Malware	0.9574	0.9709	0.9426	0.9565
Phishing	0.9770	0.9732	0.9811	0.9771

Table 3. DNS Data Performance

Task	Acc	Prec	Rec	F1
Malware	0.9518	0.9429	0.9618	0.9523
Phishing	0.9758	0.9724	0.9800	0.9762

Table 4. Geographical Data Performance

5. Conclusions

The transformer approach yields near-perfect DGA detection (F1: 0.986) with just domain names. Auxiliary data notably improve malware and phishing detection, with DNS and RDAP showing the most significant F1 gains (up to 6.2% and 4.3%, respectively). DistilBERT provides efficient and accurate featureless detection.

Acknowledgements

I want to thank my supervisor, Radek Hranický, and the whole research group for their willingness, help, and friendly approach during the solution of this work. The research is supported by the "Flow-based Encrypted Traffic Analysis" project, no. VJ02010024 granted by the Ministry of the Interior of the Czech Republic and "Smart information technology for a resilient society" project, no. FIT-S-23-8209 granted by the Brno University of Technology.

References

[1] Panpan Zhang, Tingwen Liu, Yang Zhang, Jing Ya, Jinqiao Shi, and Yubin Wang. Domain watcher: detecting malicious domains based on local and global textual features. *Procedia Computer Science*, 108:2408–2412, 2017.

[2] Sandeep Yadav, Ashwath Kumar Krishna Reddy, AL Narasimha Reddy, and Supranamaya Ranjan. Detecting algorithmically generated malicious domain names. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 48–61, 2010.

[3] Alessandro Cucchiarelli, Christian Morbidoni, Luca Spalazzi, and Marco Baldi. Algorithmically generated malicious domain names detection based on n-grams features. *Expert Systems with Applications*, 170:114551, 2021.

[4] Hong Zhao, Zhaobin Chang, Guangbin Bao, and Xiangyan Zeng. Malicious domain names detection algorithm based on n-gram. *Journal of Computer Networks and Communications*, 2019(1):4612474, 2019.

[5] Ahmad O Almashhadani, Mustafa Kaiiali, Domhnall Carlin, and Sakir Sezer. Maldomdetector: A system for detecting algorithmically generated domain names with machine learning. *Computers & Security*, 93:101787, 2020.

[6] Radek Hranický, Adam Horák, Jan Polišenský, Ondřej Ondryáš, Kamil Jeřábek, and Ondřej Ryšavý. Spotting the hook: Leveraging domain data for advanced phishing detection. In *Proceedings of the 20th International Conference on Network and Service Management (CNSM)*, pages 1–7, 2024.

[7] Paul K Mvula, Paula Branco, Guy-Vincent Jourdan, and Herna L Viktor. Covid-19 malicious domain names classification. *Expert Systems with Applications*, 204:117553, 2022.

[8] Akhila GP and Angelin Gladston. A machine learning framework for domain generating algorithm based malware detection. *Security and Privacy*, 3(6):e127, 2020.

[9] Jose Selvi, Ricardo J Rodríguez, and Emilio Soria-Olivas. Detection of algorithmically generated malicious domain names using masked n-grams. *Expert systems with applications*, 124:156–163, 2019.

[10] Congyuan Xu, Jizhong Shen, and Xin Du. Detection method of domain names generated by dgas based on semantic representation and deep neural network. *Computers & Security*, 85:77–88, 2019.

[11] Luhui Yang, Guangjie Liu, Yuewei Dai, Jinwei Wang, and Jiangtao Zhai. Detecting stealthy domain generation algorithms using heterogeneous deep neural network framework. *IEEE Access*, 8:82876–82889, 2020.

[12] Youfeng Niu, Mingxi Guan, Wenhao Yuan, Yilin Chen, Lingyi Chen, and Qiming Yu. A bayesian optimization-based lstm model for dga domain name identification approach. In *Journal of Physics: Conference Series*, volume 2303, page 012015. IOP Publishing, 2022.

[13] Ebubekir Buber, Banu Diri, and Ozgur Koray Sahingoz. Nlp based phishing attack detection from urls. In *Intelligent Systems Design and Applications: 17th International Conference on Intelligent Systems Design and Applications (ISDA 2017) held in Delhi, India, December 14-16, 2017*, pages 608–618. Springer, 2018.

[14] Luhui Yang, Guangjie Liu, Jinwei Wang, Jiangtao Zhai, and Yuewei Dai. A semantic element representation model for malicious domain name detection. *Journal of Information Security and Applications*, 66:103148, 2022.