УДК 004.02

ПОКРАЩЕННЯ КЛАСИФІКАЦІЇ ШКІДЛИВИХ URL ЗА ДОПОМОГОЮ ВЕКТОРНИХ ПРЕДСТАВЛЕНЬ НА ОСНОВІ ТРАНСФОРМЕРІВ

Глоба Л. С., Приходнюк В.В., Цуканов С.О.
Навчально-науковий інститут телекомунікаційних систем
КПІ ім. Ігоря Сікорського, Україна
E-mail: sergeyidus@gmail.com

ENHANCING MALICIOUS URL CLASSIFICATION WITH TRANSFORMER-BASED EMBEDDINGS

The basic trends in the application of transformer models for URL vectorization in malicious URL detection tasks have been summarized. The results of comparative modeling of the effectiveness of combining BERT, SBERT, and RoBERTa with neural networks (LSTM, GRU, MLP) for URL classification are presented.

Зі стрімким зростанням інтернет-трафіку у світі спостерігається значне посилення кіберзагроз, серед яких шкідливі URL-адреси (Uniform Resource Locator — Уніфікований Локатор Ресурсів) відіграють роль одного з основних векторів атак. Традиційні методи виявлення, що базуються на правилах, уже не здатні ефективно протистояти швидкій еволюції цих загроз, що підкреслює необхідність розробки більш просунутих підходів. Запропонований підхід зосереджується на аналізі сучасних технік обробки природної мови, зокрема моделей BERT (Bidirectional Encoder Representations from Transformers — Двонаправлені Кодувальні Представлення від Трансформерів), SBERT (Sentence-BERT — BERT для речень) та RoBERTa (Robustly Optimized BERT Approach — Стійкий Оптимізований Підхід BERT), у поєднанні з архітектурами нейронних мереж, такими як MLP (Multi-Layer Perceptron — Багатошаровий Персептрон), LSTM (Long Short-Term Memory — Довга Короткострокова Пам'ять), GRU (Gated Recurrent Unit — Керована Рекурентна Одиниця).
Метою роботи є не лише покращення точності класифікації, але й подолання таких викликів, як залежність від якості навчальних даних, складність моделей та високі обчислювальні витрати [8]. У цьому контексті дослідження пропонує комплексний підхід, який об’єднує передові методи векторизації та класифікації для створення надійного інструменту боротьби з кіберзагрозами.

Проблема виявлення шкідливих URL-адрес є добре дослідженою, однак сучасні методи продовжують стикатися з обмеженнями. У літературі представлено кілька підходів до вирішення цієї задачі: наприклад, один із методів пропонує виявлення фішингових URL-адрес на основі лексичних, символьних та словесних векторних ознак із використанням комбінації CNN (Convolutional Neural Network — Згорткова Нейронна Мережа) та GRU

шарів, досягаючи точності 94.4% [1], хоча відсутність порівняння з іншими сучасними методами ускладнює оцінку його ефективності; інший метод, AFSADL-MURLC [2], використовує токенізацію NLTK, вбудовування GloVe та класифікацію через GRU, досягаючи точності 98.25%, перевершуючи традиційні методи, такі як Random Forest (99.03%) чи Naive Bayes (95.37%), але поступаючись найкращим результатам; ще один підхід комбінує глибоке навчання та знання експертів для виявлення ін'єкційних атак (SQL, XSS), досягаючи точності 99.39% [4], хоча складність архітектури ускладнює практичне впровадження; гібридна модель для виявлення DDoS-атак на основі CNN та CART показала точність 97.33% [6], а модифікована версія BERT (M-BERT) [7] досягла Macro-Precision 94.42%, перевищуючи базовий BERT (91.28%), але поступаючись через гетерогенність даних. Основними викликами залишаються якість навчальних даних, складність моделей та потреба у значних обчислювальних ресурсах [8], що підкреслює необхідність нових підходів, які б поєднували високу точність із практичною реалізацією.

У цьому дослідженні класифікація URL-адрес здійснюється у два етапи: отримання векторного представлення URL та його класифікація за допомогою нейронних мереж. Для векторизації використано трансформерні моделі: BERT [10] — двонаправлена модель, яка аналізує текст у обох напрямках, забезпечуючи глибоке розуміння контексту; SBERT — покращена версія BERT для обчислення семантичної схожості між реченнями з архітектурою Siamese для швидшого генерування вбудовувань; RoBERTa — оптимізована версія BERT із динамічним маскуванням та більшим обсягом навчальних даних, що покращує узагальнюючу здатність. Векторні представлення подаються на вхід нейронних мереж: MLP із двома прихованими шарами (384 та 256 нейронів), який підходить для задач без урахування порядку даних; LSTM та GRU — рекурентні мережі, адаптовані для обробки послідовних даних, таких як текст URL-адрес.

Для оцінки ефективності підходу використано датасет "Malicious URLs dataset" [3], що містить 10,000 URL-адрес (по 5,000 шкідливих та безпечних), розділений на навчальну (80%) та тестову (20%) вибірки. Мітки класів перекодовано у числові значення: 0 для шкідливих, 1 для безпечних. Моделі реалізовано за допомогою Python та PyTorch, навчання проводилося з використанням оптимізатора Adam (learning rate 0.0001) протягом 80 епох із розміром міні-батчу 32. Для оцінки якості класифікації використано метрики: точність (accuracy), recall та precision. Експерименти показали високу ефективність запропонованих моделей: BERT + LSTM досягла найкращої точності — 99.20%, із precision 99.50% та recall 99.00% для безпечних URL-адрес і 99.60% та 99.00% для шкідливих, найкраще збалансувавши помилки першого та другого роду; BERT + GRU показала точність 99.15% із близькими показниками precision та recall; RoBERTa + LSTM досягла точності 98.65%; SBERT + GRU показала найнижчу точність — 96.30%; MLP-варіанти досягли точності до 99.10%, але поступаються рекурентним мережам у роботі з послідовними даними.

Дослідження продемонструвало, що комбінація BERT із LSTM є найбільш оптимальним рішенням для класифікації шкідливих URL-адрес, забезпечуючи точність 99.20% та високі показники precision і recall. Моделі RoBERTa + LSTM та RoBERTa + GRU також показали конкурентні результати (98.65% та 98.75% відповідно), що робить їх перспективними для подальших досліджень. Моделі на основі MLP виявилися менш ефективними для послідовних даних, хоча й досягли високих показників (до 99.10%). Запропонований підхід підтверджує переваги використання трансформерних моделей для обробки URL-адрес, дозволяючи ефективно виявляти кіберзагрози. Рекомендується подальше вдосконалення моделей із урахуванням гетерогенності даних [3,5] та оптимізації обчислювальних витрат для реального часу. Отримані результати можуть стати основою для створення практичних систем захисту від шкідливих URL-адрес у сучасних умовах зростання кіберзагроз.

Література

1. Rasymas and L. Dovydaitis, “Detection of Phishing URLs by Using Deep Learning Approach and Multiple Features Combinations,” Baltic Journal of Modern Computing, vol. 8, no. 3, Sep. 2020, doi: https://doi.org/10.22364/bjmc.2020.8.3.06
2. A. Mustafa Hilal et al., “Malicious URL Classification Using Artificial Fish Swarm Optimization and Deep Learning,” Computers, Materials & Continua, vol. 74, no. 1, pp. 607–621, 2023, doi: https://doi.org/10.32604/cmc.2023.031371
3. Malicious And Benign URLs dataset. URL: https://www.kaggle.com/datasets/siddharthkumar25/malicious-and-benign-urls (дата звернення: 20.03.2025).
4. C. Zhao, S. Si, T. Tu, Y. Shi, and S. Qin, “Deep-Learning Based Injection Attacks Detection Method for HTTP,” Mathematics, vol. 10, no. 16, p. 2914, Aug. 2022, doi: https://doi.org/10.3390/math10162914
5. CSIC 2010 Web Application Attacks dataset https://www.kaggle.com/datasets/ispangler/csic-2010-web-application-attacks (дата звернення: 20.03.2025).
6. B. Goparaju and B. S. Rao, “Distributed Denial-of-Service (DDoS) Attack Detection using 1D Convolution Neural Network (CNN) and Decision Tree Model,” Journal of Advanced Research in Applied Sciences and Engineering Technology, vol. 32, no. 2, pp. 30–41, Sep. 2023, doi: https://doi.org/10.37934/araset.32.2.3041
7. B. Yu, F. Tang, Daji Ergu, R. Zeng, B. Ma, and F. Liu, “Efficient Classification of Malicious URLs: M-BERT - A Modified BERT Variant for Enhanced Semantic Understanding,” IEEE Access, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3357095
8. F. Torregrossa, R. Allesiardo, V. Claveau, N. Kooli, and G. Gravier, “A survey on training and evaluation of word embeddings,” International Journal of Data Science and Analytics, vol. 11, no. 2, pp. 85–103, Feb. 2021, doi: https://doi.org/10.1007/s41060-021-00242-8
9. F. Torregrossa, R. Allesiardo, V. Claveau, N. Kooli, and G. Gravier, “A survey on training and evaluation of word embeddings,” International Journal of Data Science and Analytics, vol. 11, no. 2, pp. 85–103, Feb. 2021, doi: https://doi.org/10.1007/s41060-021-00242-8
10. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding,” arXiv.org, May 24, 2019. https://arxiv.org/abs/1810.04805