УДК 004.8

Д. Р. Трасіруб

Засіб для виявлення вебатак на основі технології великих мовних моделей

Вінницький національний технічний університет

Анотація

Розглянуто можливості та обґрунтовано необхідність інтеграції великих мовних моделей у системи виявлення вебатак, а також визначено основні аспекти технологічної реалізації програмного засобу.

Ключові слова: *вебатака, вебресурс, безпека, великі мовні моделі, загроза.*

Abstract

The possibilities and necessity of integrating large language models into web attack detection systems are considered and the main aspects of the technological implementation of the software tool are identified.

Keywords: *: web attack, web resource, security, large language models, threat.*

Вступ

У сучасному цифровому середовищі, де вебтехнології проникають в усі сфери суспільного життя, питання захисту вебдодатків набуває особливої актуальності. Постійне зростання кількості атак на вебресурси, складність технік зловмисників та висока вартість кіберінцидентів обумовлюють необхідність використання інтелектуальних засобів для аналізу трафіку та виявлення загроз у режимі реального часу.

Традиційні засоби, як-от WAF чи IDS/IPS, хоча й залишаються актуальними, не завжди здатні виявляти нові, обфусковані чи комбіновані атаки. У цьому контексті особливої ваги набуває використання великих мовних моделей (LLM), здатних проводити глибокий контекстуальний аналіз запитів і виявляти аномальну поведінку навіть без наявності заздалегідь визначених сигнатур..

Огляд методів і технологій штучного інтелекту

У сфері кіберзахисту вебдодатків активно застосовуються різні методи аналізу трафіку. Серед них:

- Rule-based аналіз — традиційний підхід, що базується на статичних правилах та сигнатурах. Його обмеження полягають у неспроможності адаптуватися до нових видів атак без оновлення правил.
- Машинне навчання — дозволяє класифікувати запити як нормальні чи шкідливі, використовуючи такі моделі, як дерева рішень, випадкові ліси, градієнтне бустування тощо. Використовується для аналізу логів, параметрів запиту, розміру, IP-адрес тощо.
- Нейронні мережі — забезпечують виявлення складних взаємозв’язків і нетипових шаблонів поведінки. Особливо ефективні для аналізу послідовностей, таких як історія запитів користувача
- Великі мовні моделі — здатні розуміти смислове навантаження HTTP-запитів, виявляти шкідливі інструкції, сформульовані природною мовою, та розпізнавати нестандартні форми атак.

Прогалини у сучасних дослідженнях та обґрунтування необхідності подальших розробок

Незважаючи на активне впровадження ШІ у сфері психологічної реабілітації, існує низка обмежень і прогалин, які потребують подальших досліджень. По-перше, обмежена точність алгоритмів у визначенні складних емоційних станів викликає необхідність розробки більш надійних методів аналізу психологічних даних. Існуючі алгоритми часто недостатньо добре враховують культурні, соціальні та індивідуальні відмінності, що може призводити до некоректної інтерпретації емоцій користувачів.

Другою важливою проблемою є етичні та конфіденційні аспекти. Використання чутливих даних, таких як емоційний стан або особисті переживання користувачів, ставить перед розробниками завдання забезпечення надійного захисту інформації. Існує потреба в нових підходах до зберігання, обробки та передачі даних у таких системах, що відповідають високим стандартам безпеки.

Також є обмеження у сфері адаптивності рекомендаційних систем. Більшість існуючих систем покладаються на стандартні моделі, що можуть не враховувати динамічні зміни стану користувача або специфіку різних груп населення. Подальші дослідження потрібні для розробки більш гнучких та адаптивних моделей, які можуть підлаштовуватись під змінювані умови.

Таким чином, для вдосконалення інформаційних систем психологічної підтримки з інтеграцією ШІ необхідні подальші розробки, спрямовані на підвищення точності моделей, вдосконалення конфіденційності даних та підвищення адаптивності систем.

Технологічна реалізація

Технологічна реалізація програмного засобу для виявлення вебатак побудована у вигляді проксі-сервера, що аналізує вхідні HTTP-запити та приймає рішення про їх пропуск або блокування. Основу системи становить вебфреймворк FastAPI, який дозволяє реалізувати асинхронну обробку запитів, забезпечує зручну побудову REST API та швидке розгортання сервісу. Вся логіка системи зосереджена на двох типах аналізу: rule-based і на основі великої мовної моделі.

На етапі обробки запиту сервер приймає URL, метод запиту, заголовки, IP-адресу користувача та тіло запиту, після чого формує структурований об’єкт із цими даними. У випадку, якщо активовано режим LLM, система передає цей об’єкт до зв’язки з великою мовною моделлю GPT-4o-mini через бібліотеку LangChain. Передача даних до моделі супроводжується генерацією промпту — спеціального шаблону, який включає як інструкції, так і приклади атак, критерії класифікації запиту та формат відповіді. Отримана відповідь аналізується і подається у вигляді JSON, який містить статус запиту (наприклад, "normal"або "threat"), тип можливої атаки, рівень впевненості моделі, короткий коментар та розгорнуте пояснення.

Якщо використання LLM з якихось причин неможливе (наприклад, відсутній ключ API або сталася помилка), система переходить до rule-based режиму. У цьому випадку запит аналізується за допомогою регулярних виразів, що дозволяють виявити поширені шаблони атак, такі як SQL-ін’єкції, XSS, LFI або командні ін’єкції. У разі збігу із шаблоном запит класифікується як загроза і блокується.

Усі результати класифікації, включно з повною інформацією про запит, IP-адресу користувача, джерело аналізу (LLM або rule-based) та час обробки, зберігаються у лог-файлі. Це дозволяє вести історію, виконувати подальший аналіз, а також використовувати накопичені дані для навчання чи тестування інших систем.

У випадку, якщо запит не є загрозою, він перенаправляється на тестовий уразливий вебдодаток, який працює на іншому порту. Перенаправлення реалізовано через модуль requests, а всі заголовки, окрім службових, передаються збережено. У підсумку користувач отримує відповідь безпосередньо від вебдодатку, навіть не підозрюючи про існування проміжного захисту.

Завдяки гнучкій архітектурі та модульності цей засіб може легко адаптуватися до використання з іншими LLM-моделями, змінювати набір правил або інтегруватися у більшу систему кіберзахисту. Такий підхід забезпечує баланс між інтелектуальним аналізом на основі ШІ та швидкістю rule-based методів, що є критично важливим для реального використання у системах захисту вебресурсів.

Висновок

Інтеграція великих мовних моделей у системи виявлення вебатак відкриває нові можливості для побудови сучасних, інтелектуальних засобів захисту вебресурсів. Завдяки здатності аналізувати HTTP-запити на глибшому, семантичному рівні, такі моделі дозволяють виявляти не лише відомі, а й нові або обфусковані атаки, які часто залишаються поза межами дії традиційних сигнатурних фільтрів. Поєднання LLM із rule-based підходом забезпечує надійність і гнучкість аналізу, зменшуючи ризик помилкових спрацювань та підвищуючи точність класифікації.

Реалізована система, побудована на базі FastAPI демонструє ефективну обробку запитів у реальному часі, підтримує логування, адаптивну реакцію на загрози та просту інтеграцію з іншими сервісами. Її модульна структура та можливість масштабування дозволяють розглядати цей підхід як перспективну основу для подальшого розвитку систем веббезпеки.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Аналіз вразливостей великих мовних моделей / V. Kolchenko та ін. Journal of scientific papers "social development and security". 2024. Т. 14, № 5. С. 237–246. URL: https://doi.org/10.33445/sds.2024.14.5.22 (дата звернення: 25.05.2025).
2. Дослідження загроз безпеки великих мовних моделей за допомогою автоматизованих інструментів / V. Kolchenko та ін. Journal of scientific papers "social development and security". 2024. Т. 14, № 6. С. 81–96. URL: https://doi.org/10.33445/sds.2024.14.6.9 (дата звернення: 25.05.2025).
3. Аналіз вузькоспрямованного тексту за допомогою великих мовних моделей / В. Волоховський та ін. Grail of science. 2024. № 43. С. 313–321. URL: https://doi.org/10.36074/grail-of-science.06.09.2024.041 (дата звернення: 25.05.2025).
4. Pavlenko A., Kostiuchko S. Виявлення та аналіз найвразливіших місць веб- ресурсів. Computer-integrated technologies: education, science, production. 2023.
5. № 52. С. 85–93. URL: https://doi.org/10.36910/6775-2524-0560-2023-52-11
6. (дата звернення: 25.05.2025).
7. Качан В., Марчук А. Дослідження методів захисту від витоків інформації через метадані ресурсів web-додатків. Радіоелектроніка та молодь у XXI столітті. Т. 4 : Конференція "Перспективи розвитку інфокомунікацій та інформаційно-вимірювальних технологій". Харків, Україна, 2024. URL: https://doi.org/10.30837/iyf.pdicimt.2024.066 (дата звернення: 25.05.2025)..

Трасіруб Данило Романович - студент групи 1БС-21б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: trasirubdanya@gmail.com

Науковий керівник: ***Куперштейн Леонід Михайлович*** *– к.т.н., доцент кафедри захисту інформації, Вінницький національний технічний університет, м. Вінниця email: kupershtein.lm@gmail.com*

Trasirub Danylo Romanovych - student of group 1BS-21b, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: trasirubdanya@gmail.com

Supervisor: ***Kupershtein Leonid Mikhailovich –*** *PhD, Associated Professor of Information Protection Chair, Vinnytsia National Technical University, Vinnytsia, email: kupershtein.lm@gmail.com*