НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2026 Том 26 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2026 Vol. 26 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2026-26-5-981-992
УДК 004.056

Методика формирования и набор данных HTTP-трафика для классификации веб-приложений в задачах выявления уязвимостей бизнес-логики

Алексей Алексеевич Космачев[1], Игорь Витальевич Котенко[2]✉

[1,2] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация
[1] Национальный исследовательский университет «Высшая школа экономики», Москва, 123458, Российская Федерация
[2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] a.kosmachev@hse.ru, https://orcid.org/0009-0001-7282-4784
[2] ivkote@comsec.spb.ru✉, https://orcid.org/0000-0001-6859-7120

Аннотация

Введение. Исследованы вопросы создания специализированного набора данных HyperText Transfer Protocol (HTTP) трафика для решения задачи многоклассовой классификации веб-приложений по тематическому направлению (домену) с целью последующего выявления уязвимостей бизнес-логики. Каждая запись реализованного набора данных содержит полный текст HTTP-запроса и ответа, включая заголовки, тело, путь и параметры, что отличает его от существующих аналогов, ограниченных либо одной категорией, либо уникальным адресом ресурса в интернете (Uniform Resource Locator) без содержимого. Разработанный набор данных является первым публично доступным набором данных HTTP-трафика с многодоменной разметкой по категориям, релевантным задачам обнаружения уязвимостей бизнес-логики. **Метод.** Предложенная методика автоматизированного сканирования веб-приложений включает 5 последовательных этапов: определение состава тематических категорий, выбор вспомогательных источников и метода сбора, подготовку исходного списка веб-ресурсов, многоуровневый веб-обход с сохранением полного содержимого запросов и ответов и имитацией пользовательских действий, а также постобработку и фильтрацию собранных данных. На собранных данных проведено сравнительное исследование 5 моделей классификации: трех моделей на основе Term Frequency — Inverse Document Frequency (логистическая регрессия, линейный метод опорных векторов и наивный байесовский классификатор) и двух нейросетевых архитектур (легковесная и быстрая модель FastText и сверточная нейросеть, адаптированная для работы с текстом на основе архитектуры Кима TextCNN). **Основные результаты.** В результате реализации обхода веб-приложений сформирован набор данных из 53 074 записей в бинарном колоночно-ориентированном формате для хранения и обработки больших объемов данных Parquet, охватывающий четыре категории: электронная коммерция, игровая индустрия, финансовая индустрия, социальные сети и мессенджеры. Наилучший результат сравнительного исследования достигнут моделью TextCNN. Набор данных и методология опубликованы в открытом доступе. **Обсуждение.** Ограничениями предложенного подхода к созданию набора данных HTTP-трафика для задачи многоклассовой классификации веб-приложений по тематическому направлению являются: ограничение охвата набора данных четырьмя тематическими областями, минимальный тематически значимый контент публично доступных страниц приложений категории социальных сетей и мессенджеров, внесение артефактов перевода в признаковое пространство порядка 30 % записей.

Ключевые слова

многоклассовая классификация, веб-приложения, HTTP-трафик, веб-обход, набор данных, бизнес-логика, уязвимости веб-приложений, TextCNN, TF-IDF, машинное обучение, глубокое обучение, автоматизированный сбор данных

Благодарности

Работа выполнена при частичной финансовой поддержке бюджетной темы FFZF-2025-0016.

© Космачев А.А., Котенко И.В., 2026

Ссылка для цитирования: Космачев А.А., Котенко И.В. Методика формирования и набор данных HTTP-трафика для классификации веб-приложений в задачах выявления уязвимостей бизнес-логики // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 5. С. 981–992. doi: 10.17586/2226-1494-2026-26-5-981-992

A technique and dataset of HTTP traffic for web application classification in business logic vulnerability detection

Aleksei A. Kosmachev[1], Igor V. Kotenko[2]✉

[1,2] St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation
[1] National Research University Higher School of Economics, Moscow, 123458, Russian Federation
[2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] a.kosmachev@hse.ru, https://orcid.org/0009-0001-7282-4784
[2] ivkote@comsec.spb.ru✉, https://orcid.org/0000-0001-6859-7120

Abstract
This article explores creating a specialized HyperText Transfer Protocol (HTTP) traffic dataset for multi-class web application classification by topic (domain) to identify business logic vulnerabilities. Each record contains the full text of the HTTP request and response, including headers, body, path, and parameters, which distinguishes the dataset from existing analogs limited to either a single category or a unique resource address on the Internet (Uniform Resource Locator) without any content, and is the first publicly available HTTP traffic dataset with multi-domain markup by categories relevant to the tasks of detecting business logic vulnerabilities. A technique for automated web application scanning is proposed, comprising five sequential stages: defining the composition of subject categories, selecting auxiliary sources and a collection method, preparing an initial list of web resources, multi-level web crawling with preservation of the full content of requests and responses and simulating user actions, as well as post-processing and filtering of the collected data. A comparative study of five classification models was conducted on the collected data: three models based on Term Frequency — Inverse Document Frequency (logistic regression, linear support vector machine, and naive Bayes classifier) and two neural network architectures (the lightweight and fast FastText model and a convolutional neural network adapted for the texts based on Kim's TextCNN architecture). Web application crawling resulted in a dataset of 53,074 records in Parquet format, covering four categories: e-commerce, games, finance, and social networking and messaging. The best result of model comparison was achieved by the TextCNN model. The dataset and technique are published openly. The limitations of the proposed approach to creating a dataset of HTTP traffic for the task of multi-class classification of web applications by topic are the limitation of the coverage of the dataset to four topic domains, the minimal thematically significant content of publicly accessible pages of applications in the category of social networks and instant messengers, and the introduction of translation artifacts into the feature space of about 30 % of the records.

Keywords
multi-class classification, web applications, HTTP traffic, crawling, dataset, business logic, web application vulnerabilities, TextCNN, TF-IDF, machine learning, deep learning, automated data collection

Acknowledgements
This study was supported in part by the state-funded project FFZF-2025-0016.

For citation: Kosmachev A.A., Kotenko I.V. A technique and dataset of HTTP traffic for web application classification in business logic vulnerability detection. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 5, pp. 981–992. (in Russian). doi: 10.17586/2226-1494-2026-26-5-981-992

Введение

Активное развитие технологий искусственного интеллекта, включая машинное обучение (Machine Learning, ML), способствует улучшению эффективности многих профессиональных видов деятельности человека, в том числе в области информационной безопасности. Одним из прикладных применений при этом является автоматизация моделирования атак и тестирования на проникновение [1, 2]. Интеллектуализация тестирования на проникновение способна обеспечить быстрый и глубокий анализ информационных систем, приближенный к качеству специалистов в рассматриваемой области. В настоящее время исследующие возможности ML в контексте анализа защищенности веб-приложений, демонстрируют высокие способности к повышению эффективности автоматизации выявления и эксплуатации классических уязвимостей, таких как межсайтовый скриптинг и Structured Query Language (SQL) инъекции. Однако сохраняется фундаментальная проблема обнаружения уязвимостей, требующих глубокого понимания контекста и поведения веб-приложения. Одним из классов таких уязвимостей являются уязвимости бизнес-логики.

Уязвимости бизнес-логики возникают из-за непреднамеренного наличия в функционале приложения способов нарушения заложенной логики работы, что зачастую приводит к получению определенных услуг, предоставляемых веб-приложением в обход существующих проверок. Так, например, классическим примером подобных уязвимостей является возможность множественного применения купона на скидку в интернет-магазинах, ведущее к получению существенного снижения стоимости товара, вопреки заложенной логике.

Эффективное обнаружение уязвимостей данного класса требует от исследователя безопасности глубокого понимания работы веб-приложения и высокого уровня экспертной осведомленности (способности анализировать, классифицировать и предсказывать визуальные или смысловые признаки и тренды). При этом автоматизация их обнаружения при помощи классического инструментария сильно затруднена ввиду невозможности обеспечить стандартные алгоритмы необходимыми знаниями о деталях работы каждого конкретного приложения и общем принципе обнаружения уязвимостей бизнес-логики, так как практически каждый случай выявления является уникальным. В свою очередь, алгоритмы ML обладают потенциалом к решению данной задачи.

Для алгоритмов ML требуются наборы обучающих данных, приближенные к решаемой задаче [3]. В контексте обнаружения уязвимостей бизнес-логики первичной задачей для обеспечения возможности понимания деталей работы веб-приложения является определение направленности веб-приложения, определяющей его функциональное назначение и решаемые задачи. На основе понимания бизнес-задачи приложения имеется возможность выявления общих закономерностей появления уязвимостей данного класса и более осознанного последующего изучения имеющихся в наличии функциональных возможностей.

Набор данных для рассматриваемой задачи должен удовлетворять требованиям к реализации многоклассовой классификации с возможностью определения направленностей, характерных для основных бизнес-задач современных веб-приложений. Анализ существующих наборов данных, находящихся в открытом доступе, показал отсутствие коллекций, удовлетворяющих данным критериям для решения поставленной задачи. Таким образом, целью данной работы является разработка новой методики создания набора данных и самого набора данных для решения задачи многоклассовой классификации направленности веб-приложений в контексте выявления уязвимостей бизнес-логики.

Анализ релевантных работ

Задача многоклассовой классификации направленности веб-приложений по характеристикам HyperText Transfer Protocol (HTTP) трафика в целях обнаружения уязвимостей бизнес-логики является новой и не имеет устоявшихся эталонных баз данных. Смежными областями при этом являются детектирование атак на веб-приложения, классификация сетевого трафика и автоматическая категоризация веб-сайтов.

В данных областях накоплен ряд публично доступных наборов данных, которые допускают частичное или косвенное применение в рамках поставленной задачи. Все рассматриваемые далее работы систематизированы по тому, насколько близко содержащиеся в них данные соответствуют исходной постановке: наличие HTTP-трафика с полным содержимым запросов и ответов, наличие разметки, соответствующей тематической области (домену) приложения, и широта тематического покрытия.

Классификация известных работ представлена в табл. 1.

Базы данных HTTP-трафика, ориентированные на безопасность. CSIC 2010 — один из наиболее широко цитируемых публичных наборов данных в области безопасности веб-приложений. Он содержит 36 тыс. нормальных и более 25 тыс. аномальных HTTP-запросов, автоматически сгенерированных для тестового интернет-магазина.

Каждый запрос представлен в полном объеме: метод, URI, заголовки и тело запроса. Этот набор данных размечен по двум классам (нормальный/аномальный) и включает атаки типа SQL-инъекций, межсайтового скриптинга (XSS), переполнения буфера, Carriage Return — Line Feed (CRLF) инъекций и других. Принципиальное ограничение для поставленной задачи состоит в том, что весь трафик относится к единственной тематической области приложения – электронной коммерции, а разметки по направленности не предусмотрено. Вместе с тем нормальная часть трафика (36 тыс. запросов) потенциально применима для формирования образца трафика сегмента электронной коммерции при создании многодоменного набора данных.

EClog [4] — набор данных, содержащий очищенные и анонимизированные лог-файлы реального интернет-магазина за 183 дня с суммарным объемом свыше 35 млн записей. В отличие от CSIC 2010, EClog охватывает реальный трафик продолжительного периода, что позволяет изучать сессионные и временные закономерности поведения пользователей. Вместе с тем этот набор данных не содержит информации о запросах и ответах на запросы, а включает в себя только данные о методе, URL-пути и базовых заголовках в формате стандартного лога доступа. Как и CSIC 2010, EClog ограничен одной тематической областью (электронная коммерция) и не несет разметки по направленности приложения.

Широкомасштабные источники неразмеченного HTTP-трафика. HTTP Archive — открытая база данных, периодически производящая сканирование ведущих веб-сайтов и сохраняющая результаты в формате HTTP Archive (HAR), содержащем полные заголовки и тела запросов и ответов. Данные доступны через облачное хранилище данных Google BigQuery как публичный набор данных и охватывают миллионы URL. Фундаментальное ограничение для решаемой задачи состоит в отсутствии какой-либо разметки по тематике или направленности сайтов, поэтому для применения в задаче классификации необходима внешняя категоризация, которая может быть получена, например, из иных размеченных наборов данных.

Rapid7 Project Sonar HTTP — набор данных библиотеки открытых данных для специалистов по кибербезопасности Rapid7 Open Data, содержащий ответы на стандартные HTTP/1.1 GET-запросы к публичным IP-адресам сети Интернет. Сканирования проводятся раз в две недели с 2013 года и охватывают порты 80 и 443, а также ряд нестандартных HTTP-портов; результаты публикуются в сжатом виде (gz-архивы). Особенностью данного источника является наличие только ответа сервера на запрос к главной странице; разметки по

Таблица 1. Классификация наборов данных, применимых к смежным областям
Table 1. Classification of datasets applicable to related fields

Набор данных	Тип данных	Объем	Количество категорий	Проблемы применимости к задаче
Consejo Superior de Investigaciones Cientificas(CSIC) 2010[1]	Полные HTTP-запросы	61 тыс.	одна электронная коммерция	Только одна категория, отсутствуют ответы, синтетическое веб-приложение
E-Commerce Log (EClog) [4]	HTTP-логи доступа	35 млн	одна электронная коммерция	Отсутствует полная информация о запросах и ответах, только одна категория
HTTP Archive[2]	HAR-файлы	много миллионов	отсутствует разметка	Необходимость категоризации, сложность обработки
Rapid7 Project Sonar HTTP[3]	Ответы на стандартные HTTP/1.1 GET-запросы	много миллионов	отсутствует разметка	Необходимость категоризации, недостаточная глубина анализа веб-приложений
MLDIT [5]	PCAP-файлы сетевого трафика приложений	39	9	Несоответствие категорий бизнес-функциям приложений, сложность обработки
Application Based Network Traffic Dataset[4]	PCAP-файлы сетевого трафика приложений	22	отсутствует разметка	Несоответствие категорий бизнес-функциям приложений, сложность обработки
IAB-Taxonomy Uniform Resource Locator (URL) Content Dataset[5]	URL, видимый контент HTML-страниц	35 тыс.	четырехуровневая IAB	Необходимость переопределения категорий, отсутствие HTML-структуры, неполное покрытие категорий, недостаточная глубина анализа веб-приложений
A Comprehensive Dataset for Webpage Classification [6]	URL, HTML-код страниц	1,1 млн	20	Необходимость переопределения категорий, неполное покрытие категорий, недостаточная глубина анализа веб-приложений
Dataset for Web Page Classification[6]	URL	2648	7	Отсутствует информация о запросах и ответах, неполное покрытие категорий
Curlie / DMOZ Directory[7]	URL, заголовок, редакционное описание	2,9 млн	45 тыс., иерархия	Отсутствует информация о запросах и ответах, необходимость нетривиальной перегруппировки категорий, устаревшие данные
CrowdFlower URL Categorization[8]	URL	31 тыс.	24	Отсутствует информация о запросах и ответах
Website Classification[9]	URL	1407	16	Отсутствует информация о запросах и ответах, неполное покрытие категорий
Artificial event log of an e-commerce process[10]	Журнал событий	10,5 тыс.	одна электронная коммерция	Отсутствует информация о запросах и ответах, синтетическое веб-приложение, требует значительной адаптации

[1] Giménez C.T., Villegas A.P., Marañón G.Á. HTTP data set CSIC 2010 [Электронный ресурс]. URL: https://web.archive.org/web/20130924222653/http://iec.csic.es/dataset (дата обращения: 03.04.2026).

[2] HTTP Archive [Электронный ресурс]. URL: https://httparchive.org (дата обращения: 03.04.2026).

[3] Rapid7 Project Sonar: HTTP GET Responses, Rapid7 Open Data [Электронный ресурс]. URL: https://opendata.rapid7.com/sonar.http/ (дата обращения: 03.04.2026).

[4] Application Based Network Traffic Dataset, Kaggle, 2022 [Электронный ресурс]. URL: https://www.kaggle.com/datasets/applicationdataset/applicationbasednetworktrafficdataset (дата обращения: 03.04.2026).

[5] IAB-Taxonomy URL Content Dataset, Kaggle, 2025 [Электронный ресурс]. URL: https://www.kaggle.com/datasets/ishikaaaaaathakur/iab-taxonomy-url-content-dataset (дата обращения: 03.04.2026).

[6] Janaviciute A., Liutkevicius A., Morkevicius N. Dataset for Web Page Classification, 2025 [Электронный ресурс]. URL: https://doi.org/10.5281/zenodo.15828693 (дата обращения: 03.04.2026).

[7] Curlie Directory Data, Curlie.org, OpenWebSearch.eu [Электронный ресурс]. URL: https://curlie.org/download (дата обращения: 03.04.2026).

[8] URL Categorization Dataset, CrowdFlower (Appen) [Электронный ресурс]. URL: https://data.world/crowdflower/url-categorization (дата обращения: 03.04.2026).

[9] Mehta H. Website Classification, Kaggle, 2021 [Электронный ресурс]. URL: https://www.kaggle.com/datasets/hetulmehta/website-classification (дата обращения: 03.04.2026).

[10] Alkofahi H. Artificial event log of an e-commerce process, 2023 [Электронный ресурс]. URL: https://doi.org/10.17632/csb-2scywmp.1 (дата обращения: 03.04.2026).

направленности также нет. При этом объем и широта охвата делают Rapid7 Sonar потенциально ценным вспомогательным ресурсом при условии последующей автоматической классификации.

Базы данных трафика с категоризацией по типам приложений. Набор данных Multi-Label Dataset of Internet Traffic (MLDIT) представлен в работе [5] как комплексный ресурс для классификации цифрового поведения пользователей. Набор данных охватывает трафик 39 приложений, сгруппированных в 9 категорий: социальные сети, видео, игры, коммуникации и другие. Данные представлены в формате PCAP и содержат многоуровневую разметку, в том числе по типу приложения, что концептуально близко к задаче определения направленности веб-приложения. Основное ограничение состоит в том, что трафик не относится к веб-приложениям в узком смысле, а извлечение HTTP-слоя из PCAP требует значительной предобработки; кроме того, категории приложений не совпадают напрямую с тематическими областями веб-сервисов, применимых к задачам обнаружения бизнес-логики.

Application Based Network Traffic Dataset — набор PCAP-файлов (в формате файлов для хранения перехваченного сетевого трафика) популярных приложений, опубликованный на платформе Kaggle. Набор данных охватывает 22 приложения и по большей части позволяет сопоставить захваченный трафик с тематической областью каждого приложения. Однако, как и в случае MLDIT, извлечение HTTP-уровня из PCAP и переопределение меток в терминах «направленности» веб-приложения требует существенной дополнительной работы.

Наборы данных категоризации веб-сайтов. IAB-Taxonomy URL Content Dataset — крупномасштабный набор данных, собранный с использованием стандарта категоризации контента Interactive Advertising Bureau (IAB). Содержит более 35 тыс. веб-страниц, размеченных по четырем уровням иерархии IAB, от широких категорий первого уровня («Arts & Entertainment», «Finance») до детальных подкатегорий четвертого уровня. Для каждого URL хранятся заголовок, описание, ключевые слова и текст страницы, извлеченные с помощью Selenium и BeautifulSoup. Использование стандарта IAB является существенным преимуществом, поскольку категории IAB частично пересекаются с тематическими областями, важными для тестирования безопасности (финансы, медицина, электронная коммерция и др.). Тем не менее для задачи выявления уязвимостей бизнес-логики иерархия IAB охватывает не все значимые прикладные тематические области и требует переопределения классов.

A Comprehensive Dataset for Webpage Classification [6] — один из крупнейших публично доступных наборов данных для классификации веб-страниц, содержащий свыше 1 069 715 URL с разметкой по трем основным меткам (Malicious, Benign, Adult), т. е. меткам (вредоносный, безопасный, для взрослых) и 20 детальным подкатегориям для легитимного контента. Данные хранятся в формате JSONL, каждая запись включает URL, метку, подметку, HTTP-статус и HTML-содержимое страницы. Присутствие HTML-контента делает этот набор данных особенно ценным: из него можно извлечь HTTP-признаки без повторного веб-обхода. Ограничением является то, что 20 подкатегорий ориентированы прежде всего на классификацию для задач веб-фильтрации (News, Society, Arts, Sports, Kids, Spam и т. п.), а не на тематические области, характерные для тестирования бизнес-логики.

Dataset for Web Page Classification — компактный набор данных из 2648 URL с разметкой по 7 тематическим категориям: business, education, adult, games, health, sport, travel. Данные представлены только в виде пар URL — категория без HTML-содержимого, что требует самостоятельного веб-обхода. Категориальный набор отличается малым числом классов и не покрывает ряд значимых для тестирования безопасности тематических областей.

Curlie/DMOZ Directory — крупнейший каталог веб-сайтов, полностью составленный и редактируемый вручную, доступный для загрузки под открытой лицензией. Данные Curlie охватывают около 2,9 млн записей, организованных примерно в 45 тыс. иерархических категорий. Формат TSV содержит URL, заголовок и редакционное описание сайта. Каталог является ценным источником размеченных URL с чрезвычайно широким тематическим охватом, однако для использования в задаче классификации трафика необходим полный веб-обход сайтов для получения HTTP-данных, а иерархическая структура категорий требует нетривиальной перегруппировки под задачу распознавания прикладной тематической области.

CrowdFlower URL Categorization — набор данных объемом более 31 тыс. URL с разметкой по 24 тематическим категориям, основанным на принципах поисковой классификации (Finance, Health, Shopping, Travel, Adult и др.). Как и в случае Curlie, данные содержат только URL и категорию без сохраненного содержимого страниц, что требует дополнительного веб-обхода.

Website Classification — небольшой набор данных из 1407 записей, содержащий URL, тематическую категорию и очищенный текст страницы для 16 классов. Несмотря на малый объем, набор данных примечателен тем, что для каждой записи уже имеется извлеченный текстовый контент, что частично компенсирует отсутствие HTTP-трафика. По своей структуре он наиболее близок к формату, необходимому для задачи категоризации веб-приложений, однако требует существенного расширения как по объему, так и по составу признаков.

Синтетические данные и данные событийных журналов. Artificial event log of an e-commerce process — синтетический набор данных, опубликованный на платформе Mendeley. Он включает данные, содержащие журналы событий, сгенерированные с помощью специализированной среды моделирования на основе модели Петри CPN Tools с использованием системы управления контентом OsCommerce (версия 2.3.4.1). Набор данных включает 910 смоделированных сессий (около 10,5 тыс. событий из 56 классов) с ролевым разграничением между администратором, гостем и зарегистрированным пользователем. Хотя данный ресурс формально описывает поведение пользователей

приложения e-commerce, в нем отсутствуют как пути HTTP-запросов, так и тела сообщений; фактически он представляет собой абстрактное логирование бизнес-процесса, а не сетевой трафик. По этой причине применение этого набора данных к задаче классификации направленности веб-приложений требует значительной адаптации.

Таким образом, анализ существующих наборов данных показывает, что ни один из рассмотренных ресурсов не обеспечивает одновременного выполнения трех ключевых требований: наличие полного HTTP-трафика (запрос и ответ с заголовками и телом); разметка по тематическим категориям, релевантным задачам выявления уязвимостей бизнес-логики; многокатегориальный охват.

Это обусловливает актуальность создания нового набора данных, предлагаемого в настоящей работе.

Методика создания набора данных

Предлагаемая методика отличается от существующих подходов к формированию наборов данных HTTP-трафика по трем ключевым аспектам:

1) многодоменная разметка по тематическим категориям, характерным для уязвимостей бизнес-логики, в отличие от однодоменных наборов данных (CSIC 2010, EClog);
2) сохранение полного содержимого HTTP-запроса и ответа: заголовков, тела, пути и параметров, в отличие от наборов данных на основе URL или журналов доступа;
3) рекурсивный обход гиперссылок нескольких уровней, обеспечивающий представление функциональных страниц приложения, а не только его главной страницы.

В работе были определены бизнес-направленности веб-приложений, в рамках которых будет производиться дальнейшая классификация. Выбрано четыре класса: электронная коммерция (e-commerce); игровая индустрия (games); финансовая индустрия (finance); социальные сети и мессенджеры (social network and messaging). Указанное количество категорий определено их достаточностью для демонстрации применимости рассматриваемой методологии к решению задачи исследования. Выбор самих направленностей обусловлен наличием у данных категорий характерных и ярко выраженных уязвимостей бизнес-логики, присущих исключительно им. При этом данные категории являются одними из самых исследованных в контексте уязвимостей бизнес-логики, популярных среди пользователей, а также обладающими достаточной близостью между собой в определенных аспектах, создавая тем самым вызов для алгоритмов ML.

Используемая в работе методика создания набора данных включает основные этапы, представленные на рис. 1.

Рассмотрим подробнее этапы методики (рис. 1).

Этап 1. Определение количества и состава категорий. Вход: требования к задаче многоклассовой классификации направленности веб-приложений в контексте выявления уязвимостей бизнес-логики. Преобразование: анализ предметной области и отбор тематических категорий, обладающих характерными и ярко выраженными уязвимостями бизнес-логики, достаточной взаимной близостью для создания нетривиальной задачи классификации и широкой представленностью среди современных веб-приложений. Выход: четыре целевые категории e-commerce; games; finance; social network and messaging.

Этап 2. Определение метода сбора и вспомогательных наборов данных. Вход: четыре целевые категории; анализ существующих открытых наборов данных HTTP-трафика и категоризации веб-сайтов (табл. 1). Преобразование: выбор открытых размеченных наборов данных URL-адресов, наиболее полно покрывающих целевые категории (CrowdFlower URL Categorization, Website Classification); обоснование метода автоматизированного многоуровневого обхода веб-приложений как основного способа сбора HTTP-трафика. Выход: определенные источники и параметры сбора данных.

Этап 3. Подготовка данных. Вход: исходные URL-адреса из выбранных наборов данных. Преобразование: отбор записей, соответствующих целевым категориям; унификация форматов; ручное дополнение категории social network and messaging; нормализация URL-адресов; удаление дублирующихся доменных имен; проверка актуальности Domain Name System (DNS) записей и доступности ресурсов по протоколам HTTP/HTTPS. Выход: список 1325 доступных уникальных URL с тематической разметкой.

Этап 4. Сбор данных. Вход: список 1325 URL; параметры обхода — глубина рекурсии, ограничения на число ссылок на каждом уровне. Преобразование: последовательный обход каждого веб-приложения на три уровня глубины с сохранением полного HTTP-трафика (заголовки, тело запроса и ответа); имитация POST-запросов посредством автоматического заполнения HTML-форм; нормализация языка контента с применением машинного перевода неанглоязычных страниц с сохранением структуры HTML-разметки. В целях обеспечения баланса между информативностью собираемых данных и вычислительными затратами на их сбор был выбран уровень рекурсии, равный трем: главная страница веб-приложения, как правило, содержит преимущественно навигационный и рекламный



Рис. 1. Последовательность основных этапов предлагаемой методики создания набора данных

Fig. 1. Main stages of the proposed methodology for creating the dataset

контенты, тогда как семантически значимые признаки, характерные для конкретной тематической категории, сосредоточены на страницах более глубоких уровней; экспериментальное подтверждение. Выход: необработанный массив 53 074 записей HTTP-трафика.

Этап 5. Анализ и постобработка полученных данных. Вход: необработанный массив 53 074 записей. Преобразование: автоматическое удаление записей, не несущих диагностически значимой информации о тематической принадлежности приложения; ручная верификация неоднозначных случаев; обучение и оценка классификационных моделей различных архитектур; анализ влияния глубины обхода на качество классификации. Выход: очищенный набор данных в формате Parquet; оценка качества набора данных и рекомендации по выбору оптимальных параметров методики.

Предполагается, что наибольшей эффективности при определении бизнес-направленности приложения методом черного ящика алгоритмы ML смогут достичь при наличии достаточной информации о возможных коммуникациях между пользователем и приложением. При этом одной из наиболее важных составляющих является полный текст ответа самого приложения, включающий присутствующие HTML-конструкции, которые могут быть характерны для конкретной направленности веб-приложения. По указанной причине, для решения поставленной задачи, обучающий набор данных должен обладать этой информацией. Анализ существующих наборов данных показал их недостаточность в контексте исследования. Тем не менее, некоторые из них могут быть использованы, но требуют определенной обработки или дальнейших действий. Следовательно, для настоящего исследования в качестве основы были выбраны наборы данных CrowdFlower URL Categorization и Website Classification, содержащие размеченные данные, подходящие к рассматриваемым по категориям.

Для возможности совместного использования двух разных наборов данных в контексте решаемой задачи были отобраны ссылки, удовлетворяющие ранее определенным категориям и унифицированы форматы представления данных. Полученный при этом набор данных обладал небольшим дисбалансом классов с перевесом категорий, исключающих социальные сети. С учетом этого, далее были вручную добавлены несколько актуальных ресурсов, заведомо относящихся к данной категории и являющихся популярными среди пользователей сети Интернет. Из-за возможности наличия устаревших ресурсов была произведена проверка существования веб-приложений, актуализован их целевой протокол коммуникации прикладного уровня (HTTP/HTTPs), а также произведена отчистка от повторяющихся значений.

Для обогащения разрабатываемого набора данных информацией о данных, возвращаемых при ответе на запросы к веб-приложению, был применен веб-обход имеющихся к текущему этапу ссылок. При этом, важным аспектом работы также является исследование связи между количеством возможных коммуникаций с целевым веб-приложением и качеством определения направленности алгоритмами ML. В результате, кроме простого сбора информации, полученной при посещении главной страницы приложения, производился рекурсивный обход веб-страниц, на которые ссылаются уже посещенные страницы, а также выполнялось заполнение имеющихся форм. Данный подход позволяет получить более полную информацию о доступной функциональности веб-приложения, а также дает возможность имитировать действия реального пользователя. В целях соблюдения равновесия между скоростью и качеством был выбран уровень рекурсии, равный трем. В процессе веб-обхода производилось сохранение только успешно завершившихся запросов. Также выполнялось определение языка веб-страницы и перевод при необходимости на единый язык — в настоящей работе — английский язык. Данное действие производилось в целях сохранения единого языка во всем наборе данных для более качественного последующего обучения моделей.

По результатам веб-обхода был составлен набор данных (табл. 2), содержащий следующие поля: категория ресурса; базовый URL; путь запроса; глубина запроса (первый, второй или третий в зависимости от уровня рекурсии); GET-параметры запроса (при их наличии); тело POST-запроса (при наличии); код ответа; заголовки ответа; тело ответа; отметка о произведении машинного перевода страницы. Следует отметить, что в некоторых полях приведенных примеров представлено только начало соответствующего значения из-за его больших размеров.

Полученные данные являются основной информацией, которую возможно получить при помощи коммуникации с веб-приложением по протоколу HTTP. Также данная информация является потенциально полезной с точки зрения дальнейшего обучения моделей. Сбор набора данных производился в процессе веб-обхода.

За счет возможного наличия дефектов веб-обхода, а также проблем с целевыми ресурсами, которые не были выявлены ранее (например, защита от обхода, геополитические ограничения, доменные имена с истекшим сроком действия), выполнена очистка набора данных. Для этого произведено удаление наиболее характерных вхождений нерелевантных данных на основе анализа сохраненного тела ответа веб-приложения, а также осуществлена пометка вхождений, обладающих косвенными признаками неподходящих данных, после чего выполнена ручная проверка отмеченных таким образом вхождений.

В результате составлен набор данных, содержащих данные о возможной HTTP-коммуникации с различными веб-приложениями, относящимися к четырем категориям выбранным на этапе 1 методики. При этом разработанный набор данных содержит имитацию действий пользователя, а также информацию об уровне рекурсии сбора данных.

Далее была проведена серия экспериментов с использованием разработанного набора данных. В ходе экспериментов произведено обучение классических моделей ML и глубокого обучения (Deep Learning, DL), используемых в задачах классификации текстовых массивов данных. Для дальнейшего анализа результатов в ходе экспериментов выполнен сбор стандартных для ML-метрик: точность (precision), полнота (recall),

Таблица 2. Пример данных, полученных в результате веб-обхода
Table 2. Example of the data obtained from web crawling

Параметр	Пример 1	Пример 2	Пример 3
Категория ресурса	Finance	Games	Social Networking and Messaging
Базовый URL	https://fiatc.es/	https://soccerproject.com	https://www.chat27.co.za/
Путь запроса	/noticias	/login.php	/
Глубина запроса	3	3	1
GET-параметры запроса	if any	if any	if any
Тело POST-запроса	—	{“section”: “4”, “step”: “”, “uID”: “0”, “eID”…	—
Код ответа	200	200	200
Заголовки ответа	{“Date”: “Sun, 15 Mar 2026 23:18:49 GMT”, “Con…	{“Server”: ”nginx/1.18.0 (Ubuntu)”, “Date”: “S…	{“Content-Type”: “text/html”, “Transfer-Encodi…
Тело ответа	html<html class=”ltr” dir=”ltr” lang=”es-ES”>\...	<!DOCTYPE htrml PUBLIC “-//W3C//DTD XHTML 1.0 …	<html lang=en>\n <head>\n <title>Chat27</t…
Отметка о произведении машинного перевода	True	False	False

доля правильных ответов (accuracy), макроусредненная F-мера (macro-averaged F1).

Совокупность описанных технических решений, таких как рекурсивный веб-обход с имитацией POST-запросов, автоматический перевод с сохранением HTML-структуры и специализированная разметка по бизнес-ориентированным тематическим категориям, образует методику, не имеющую прямого аналога среди опубликованных наборов данных HTTP-трафика (табл. 1).

Краткое описание реализации

Практическая реализация методики выполнена программными средствами на языке Python и включает четыре ключевых компонента. Формирование исходного списка URL произведено путем отбора и унификации записей из наборов данных CrowdFlower URL Categorization и Website Classification, дополненных вручную ресурсами категории социальных сетей; исходный список составил 2742 записи. Предобработка списка включала нормализацию URL, устранение дублирующихся доменных имен, проверку актуальности DNS-записей и HTTP/HTTPS-доступности ресурсов, в результате которой был сформирован итоговый список из 1325 доступных URL первичной обработки данных: для категории e-commerce количество URL 429; games — 411; finance — 354; social networking and messaging — 131.

Веб-обход реализован как асинхронный трехуровневый агент на основе Python-библиотеки aiohttp, выполняющий обход методом черного ящика с ограничением на число посещаемых ссылок на каждом уровне (до 25 на глубине второго уровня и до 50 на глубине третьего уровня) и имитацией пользовательских действий посредством автоматического заполнения и отправки HTML-форм. Для сохранения единого языкового представления данных реализован модуль автоматического перевода неанглоязычного контента с сохранением HTML-структуры разметки. Собранные записи прошли автоматическую фильтрацию (удаление страниц с кодами ошибок, признаками выставленных на продажу доменов и блокировки обхода) и ручную верификацию неоднозначных случаев, после чего сохранены в формате Parquet.

Характеристики итогового набора данных. По итогам обхода, до применения финальной фильтрации, было собрано 53 074 записи от 1325 веб-приложений. Среднее число записей на одно приложение составило около 40. Распределение по категориям после этапа веб-обхода: для e-commerce количество записей 18 500; games — 15 729; finance — 13 865; social networking and messaging — 4980.

Дисбаланс классов в итоговом наборе данных объясняется как различием в исходном количестве URL по категориям, так и различной возможностью к веб-обходу ресурсов: финансовые сайты, как правило, имеют большее число внутренних страниц с открытым доступом, тогда как социальные сети в значительной мере ограничивают доступ к контенту для неаутентифицированного пользователя. Распределение по глубине обхода следующее: для глубины первого уровня количество записей 1363; для второго — 18 433; для третьего — 33 278.

Преобладание записей третьего уровня (62,7 % от общего числа) обусловлено разветвленностью структуры большинства веб-приложений.

Среди выполненных запросов 76,8 % составляют GET-запросы (40 787 записей) и 23,2 % – POST-запросы (12 287 записей). В наборе данных доля записей без автоматического перевода составляет 69,7 % (36 997 записей), в то время как переведенные записи — 30,3 % (16 077 записей). Подавляющее большинство записей (53 071) соответствуют коду ответа 200 (OK); единич-

ные записи содержат коды 202, 302 и 307, что свидетельствует о высоком уровне отклика целевых ресурсов.

Экспериментальная оценка набора данных

Постановка задачи и методика оценки. Цель экспериментальной части — верификация набора данных посредством обучения и оценки ряда классификационных моделей на задаче предсказания тематической направленности веб-приложения по его HTTP-трафику. Для обеспечения корректности оценки обобщающей способности моделей разбиение на обучающую и тестовую выборки выполнялось с учетом принадлежности записей к отдельным веб-сайтам: использовался GroupShuffleSplit (группировка по полю base_url, test_size = 0,2, random_state = 42), что гарантирует отсутствие пересечения ресурсов между выборками (Site overlap = 0). Итоговый размер обучающей выборки составил 41 898 записей, тестовой — 11 120 записей.

В результате дисбаланса классов во всех моделях применялась компенсация через взвешивание: веса классов вычислялись методом balanced библиотеки scikit-learn. В качестве основной метрики оценки использовалась macro-averaged F1, чувствительная к предсказаниям как мажоритарных, так и миноритарных классов, а в качестве дополнительной метрики применялась accuracy.

Конструирование признакового пространства (для моделей ML). Для традиционных моделей ML была реализована схема конкатенации нескольких пространств Term Frequency — Inverse Document Frequency (TF-IDF), кодирующих различные аспекты HTTP-трафика (табл. 3).

Итоговая матрица признаков формировалась горизонтальной конкатенацией (scipy.sparse.hstack) и имела размерность 53 018 × 60 767.

Модели традиционного ML. В качестве базовых классификаторов были использованы три модели: логистическая регрессия (LogisticRegression); линейный метод опорных векторов (LinearSVC); наивный байесовский классификатор (Complement Naive Bayes, CNB).

Логистическая регрессия (параметр, обратный силе регуляризации (C) = 5,0, solver=lbfgs, multi_class=multinomial, class_weight=balanced) показала macro-averaged F1 равную 0,6531 при accuracy 0,6867. Наиболее точно классифицируется категория finance (macro-averaged F1 равная 0,81), наибольшую сложность представляет категория social networking and messaging (macro-averaged F1 равная 0,53), что объясняется как малым объемом этой категории в тестовой выборке, так и низким recall (0,41), — вероятно, из-за недостаточного текстового содержимого страниц, требующих аутентификации.

LinearSVC (C=1,0, class_weight=balanced, обернут в CalibratedClassifierCV) показал macro-averaged F1 равную 0,6263 при accuracy 0,6744. Особенностью этой модели является высокая precision по категории social networking and messaging (0,92), однако очень низкий recall (0,31), что приводит к итоговому macro-averaged F1 равному 0,46, что значительно хуже, чем у логистической регрессии.

CNB (параметр сглаживания Лапласа (alpha) = 0,1, признаки предварительно масштабированы через MaxAbsScaler, веса классов учтены через sample_weight) продемонстрировал лучший результат среди ML-моделей: macro-averaged F1 0,7009 при accuracy 0,7213. Отличительной чертой является более сбалансированный recall по всем классам, включая social networking and messaging (recall 0,57 против 0,41 у логистической регресии).

Нейросетевые модели. Для проверки способности нейросетевых архитектур извлекать семантические конструкции из тела HTTP-ответа (HTML-содержимое) была реализована серия моделей на PyTorch. В качестве входных данных использовалось только поле response_body после удаления HTML-тегов <script> и <style> и приведения к нижнему регистру. Токенизация (разбиение текста на единицы-токены) производилась на уровне слов (символы [a-z0-9]); словарь строился по обучающей части выборки с ограничением max_features = 30 000 и минимальной частотой min_df = 3. Максимальная длина последовательности — 256 токенов. Размерность векторных вложений (эмбеддингов) — 128. Все модели обучались оптимизатором AdamW (lr=10–3, weight_decay=10–4) с планировщиком ReduceLROnPlateau (patience=2, factor=0,5) и ранней остановкой при отсутствии улучшения macro-averaged F1 на тестовой выборке в течение трех эпох.

FastText (усреднение эмбеддингов через EmbeddingBag и двухслойный многослойный перцептрон (Multilayer Perceptron, MLP) со случайным отключением нейронов (dropout) показал macro-averaged F1 0,6560 при accuracy 0,6700. Модель сходится быстро, однако демонстрирует слабую различимость между категориями games и e-commerce.

TextCNN (параллельные свертки с ядрами размеров 2, 3, 4 по 128 фильтров каждое и глобальный max-

Таблица 3. Схема конкатенации TF-IDF-пространств
Table 3. TF-IDF space concatenation scheme

Признаковый блок	Трансформер	Параметры
Тело ответа (response_body)	TfidfVectorizer	max_features=50 000, sublinear_tf, *n*-gram (1,2), min_df=3
URL-путь (path)	TfidfVectorizer (char-wb)	max_features=10 000, *n*-gram (1,3), min_df=2
Сервер (Server-заголовок)	TfidfVectorizer (char-wb)	max_features=2 000, *n*-gram (1,2), min_df=2
Content-Type	TfidfVectorizer (char-wb)	max_features=1 000, *n*-gram (1,2), min_df=2
Числовые признаки	—	статусы 2xx/3xx/4xx, depth, is_translated, has_post

pooling с применением MLP) показал лучший результат среди нейросетевых моделей: macro-averaged F1 0,7456 при accuracy 0,7640. Высокие значения F1 по finance (0,85) и games (0,75) свидетельствуют об эффективности локально-контекстных *n*-граммовых представлений для данной задачи.

Итоговое сравнение моделей. Сводная таблица результатов всех рассматриваемых моделей на полном наборе данных представлена в табл. 4.

Нейросетевые архитектуры с контекстным кодированием (TextCNN) стабильно превосходят традиционные методы ML на основе TF-IDF, демонстрируя прирост macro-averaged F1 порядка 4–5 процентных пунктов относительно лучшего базового метода (CNB). Наибольшую сложность для всех моделей представляет категория social networking and messaging из-за ограниченного объема и специфики контента страниц данной категории. Категория finance классифицируется наиболее точно благодаря высокой терминологической специфичности финансового контента.

Анализ результатов

Наилучшим результатом в проведенных экспериментах является macro-averaged F1 равная 0,746 (accuracy 0,764), достигнутый моделью TextCNN. Среди базовых методов ML лидирует CNB с macro-averaged F1 равной 0,701 при accuracy равной 0,721, что на 7–10 процентных пунктов выше, чем у Logistic Regression (0,653) и LinearSVC (0,626). Прирост TextCNN относительно лучшего базового метода ML составляет приблизительно 4,5 процентных пункта по macro-averaged F1, что свидетельствует об эффективности контекстно-зависимых сверточных признаков при классификации на основе тела HTTP-ответа.

Распределение ошибок по классам обнаруживает характерную асимметрию. Категория finance демонстрирует наиболее высокие значения precision и recall во всех моделях (macro-averaged F1 равно 0,81–0,85), что объясняется высокой терминологической специфичностью финансового веб-контента: специализированная лексика (котировки, тикеры, условия вкладов, биржевые термины) практически не пересекается с тематическими полями других категорий. Категория games также классифицируется достаточно уверенно (macro-averaged F1 равно 0,66–0,75), несмотря на широкий спектр жанров — от браузерных игр до крупных игровых платформ. Наибольшую сложность представляет категория social networking and messaging: recall по ней во всех моделях остается невысоким (0,41–0,57 в ML-моделях, 0,44 у TextCNN). Основная причина — структурная особенность данного класса веб-приложений: их публично доступные страницы содержат минимум текстового контента (формы входа, маркетинговые слоганы, заглушки для неаутентифицированных пользователей), тогда как содержательный контент скрыт за авторизацией. Отдельного внимания заслуживает поведение LinearSVC: модель демонстрирует аномально высокую precision по классу social networking and messaging (0,92), однако крайне низкий recall (0,31), что свидетельствует о том, что модель «боится» ошибочно предсказать этот класс и предпочитает не предсказывать его вовсе.

Таблица 4. Сводная таблица результатов всех рассматриваемых моделей на полном наборе данных

Table 4. Summary table of the results for all models evaluated on the full dataset

Модель	Тип	Macro-averaged F1	Accuracy
TextCNN	DL	0,746	0,764
CNB	ML	0,701	0,721
FastText	DL	0,656	0,670

Неожиданно высокий результат CNB по сравнению с Logistic Regression и LinearSVC объясняется тем, что CNB изначально спроектирован для несбалансированных задач классификации текста: при оценке вероятности принадлежности к классу он опирается на статистику «дополнения», т. е. учитывает, насколько слово типично для всех прочих классов, что делает его менее чувствительным к доминирующим классам.

Каждая запись реализованного набора данных содержит полный текст HTTP-запроса и ответа, включая заголовки, тело, путь и параметры, что отличает набор данных от существующих аналогов, ограниченных либо одной категорией, либо уникальным адресом ресурса в интернете URL без содержимого, и является первым публично доступным набором данных HTTP-трафика с многодоменной разметкой по категориям, релевантным задачам обнаружения уязвимостей бизнес-логики. Полученный в результате исследования набор данных доступен публично[1].

Ограничения и направления дальнейших исследований

Ряд факторов ограничивает обобщаемость полученных результатов. Во-первых, набор данных охватывает только четыре тематических области, что снижает применимость выводов к более широкому спектру веб-приложений. Расширение числа категорий является приоритетным направлением. Во-вторых, значительная доля контента категории social networking and messaging собрана с неаутентифицированных версий страниц, что систематически снижает информативность признаков для данного класса; потенциальным решением является включение результатов браузерной эмуляции с поддержкой JavaScript. В-третьих, около 30 % записей набора данных прошли машинный перевод, что вносит артефакты перевода в признаковое пространство.

Перспективными направлениями являются: применение предобученных трансформерных моделей (Bidirectional Encoder Representations from Transformers, Robustly Optimized BERT Approach) [7] — с учетом ограничения максимальной длины последовательности и высокого уровня шума во входных данных это

[1] WebApp-Domain-Classification-Crawled-Dataset-v1, Github [Электронный ресурс]. URL: https://github.com/TheMostKnown/WebApp-Domain-Classification-Crawled-Dataset/releases/tag/v1 (дата обращения: 17.05.2026).

нетривиальная задача; использование мультимодального подхода, объединяющего текстовые признаки тела ответа с URL-признаками и структурными признаками древовидной структуры HTML-документа; а также исследование методов активного обучения для сокращения стоимости разметки при масштабировании набора данных на новые категории.

Заключение

В работе предложена и реализована методика формирования набора данных HTTP-трафика, ориентированного на задачу тематической классификации веб-приложений. Разработанная методика является первой специализированной методикой для данного класса задач, что подтверждается результатами анализа известных научных работ (табл. 1). Сформированный набор данных объемом 53 074 записей по четырем категориям (e-commerce, games, finance, social networking and messaging) получен посредством автоматизированного обхода 1325 веб-приложений с обходом гиперссылок до глубины трех уровней; каждая запись содержит заголовки и тело как запроса, так и ответа, что принципиально отличает его от преобладающих в открытом доступе наборов данных на основе только URL или HTML главной страницы. Полученный в результате исследования набор данных доступен публично.

Выполненное сравнение моделей показало, что нейросетевая архитектура TextCNN превосходит традиционные модели машинного обучения, достигая macro-averaged F1 равную 0,746 при accuracy равной 0,764, тогда как лучший из ML-методов — Complement Naive Bayes — обеспечивает macro-averaged F1 равную 0,701. Полученные результаты сопоставимы с показателями аналогичных работ по классификации веб-контента и в ряде случаев превышают их, принимая во внимание более высокий уровень шума входных данных и несбалансированность классов.

Исследование по глубине обхода установило, что переход от одноуровневого обхода к двухуровневому дает прирост macro-averaged F1 на 14 процентных пунктов (с 0,50 до 0,64), тогда как дальнейшее углубление до третьего уровня значительно не улучшает метрику при трехкратном увеличении объема данных. Данный результат подтверждает практическую целесообразность ограничения обхода двумя уровнями вложенности как компромисса между информативностью признакового пространства и вычислительными затратами на сбор данных.

Среди выявленных ограничений необходимо выделить три ключевых. Во-первых, охват набора данных ограничен четырьмя тематическими областями, что сужает применимость обученных моделей к более широкому спектру веб-приложений. Во-вторых, публично доступные страницы приложений категории social networking and messaging содержат минимальный тематически значимый контент ввиду необходимости аутентификации, что систематически снижает recall по данному классу (0,44 для TextCNN). В-третьих, порядка 30 % записей подверглись машинному переводу, что вносит артефакты перевода в признаковое пространство.

Перспективными направлениями дальнейших исследований являются: расширение числа тематических категорий набора данных; применение предобученных трансформерных моделей с учетом ограничений максимальной длины последовательности и высокого уровня контентного шума; разработка мультимодального подхода, объединяющего текстовые признаки тела ответа с URL-признаками и структурными признаками древовидной структуры HTML-документа; а также использование методов активного обучения для снижения стоимости разметки при масштабировании набора данных на новые категории. Отдельный интерес представляет исследование применимости разработанного классификатора в составе автоматизированных систем тестирования безопасности для приоритизации проверок бизнес-логики в зависимости от определенной тематической области приложения.

В следующей работе авторов будут рассмотрены детали реализации отдельных этапов работы с данными представленного набора, экспериментальное подтверждение и обоснование выбора оптимальной глубины веб-обхода, а также будет представлен детальный анализ влияния глубины обхода на информативность собранных данных.

Литература

1. Kotenko I., Chechulin A. Computer attack modeling and security evaluation based on attack graphs // Proc. of the IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS). 2013. V. 2. P. 614–619. doi: 10.1109/IDAACS.2013.6662998
2. Kotenko I., Chechulin A., Novikova E. Attack modelling and security evaluation for Security Information and Event Management // Proc. of the International Conference on Security and Cryptography. 2012. P. 391–394. doi: 10.5220/0004063403910394
3. Tushkanova O., Levshun D., Branitskiy A., Fedorchenko E., Novikova E., Kotenko I. Detection of cyberattacks and anomalies in cyber-physical systems: approaches, data sources, evaluation // Algorithms. 2023. V. 16. N 2. P. 85. doi: 10.3390/a16020085
4. Chodak, G., Suchacka, G., Chawla, Y. HTTP-level e-commerce data based on server access logs for an online store // Computer Networks. 2020. V. 183. P. 107589. doi: 10.1016/j.comnet.2020.107589

References

1. Kotenko I., Chechulin A. Computer attack modeling and security evaluation based on attack graphs. *Proc. of the IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, 2013, vol. 2, pp. 614–619. doi: 10.1109/IDAACS.2013.6662998
2. Kotenko I., Chechulin A., Novikova E. Attack modelling and security evaluation for Security Information and Event Management. *Proc. of the International Conference on Security and Cryptography*, 2012, pp. 391–394. doi: 10.5220/0004063403910394
3. Tushkanova O., Levshun D., Branitskiy A., Fedorchenko E., Novikova E., Kotenko I. Detection of cyberattacks and anomalies in cyber-physical systems: approaches, data sources, evaluation. *Algorithms*, 2023, vol. 16, no. 2, pp. 85. doi: 10.3390/a16020085
4. Chodak, G., Suchacka, G., Chawla, Y. HTTP-level e-commerce data based on server access logs for an online store. *Computer Networks*, 2020, vol. 183, pp. 107589. doi: 10.1016/j.comnet.2020.107589

5. Li W., Quenard G. Towards a multi-label dataset of internet traffic for digital behavior classification // Proc. of the 3rd International Conference on Computer Communication and the Internet (ICCCI). 2021. P. 38-46. doi: 10.1109/ICCCI51764.2021.9486831
6. Al-Maamari M., Istaiti M., Zerhoudi S., Dinzinger M., Granitzer M., Mitrovi J. A comprehensive dataset for webpage classification // International Open Search Symposium. 2024. doi: 10.5281/zenodo.10775260
7. Muhetaer M., Meng X., Zhu J., Aikebaier A., Zu L., Bai Y. Symmetry and asymmetry in pre-trained transformer models: a comparative study of TinyBERT, BERT, and RoBERTa for Chinese educational text classification // Symmetry. 2025. V. 17. N 11. P. 1812. doi: 10.3390/sym17111812

5. Li W., Quenard G. Towards a multi-label dataset of internet traffic for digital behavior classification. *Proc. of the 3rd International Conference on Computer Communication and the Internet (ICCCI)*, 2021, pp. 38-46. doi: 10.1109/ICCCI51764.2021.9486831
6. Al-Maamari M., Istaiti M., Zerhoudi S., Dinzinger M., Granitzer M., Mitrovi J. A comprehensive dataset for webpage classification. *International Open Search Symposium*, 2024. doi: 10.5281/zenodo.10775260
7. Muhetaer M., Meng X., Zhu J., Aikebaier A., Zu L., Bai Y. Symmetry and asymmetry in pre-trained transformer models: a comparative study of TinyBERT, BERT, and RoBERTa for Chinese educational text classification. *Symmetry*, 2025, vol. 17, no. 11, pp. 1812. doi: 10.3390/sym17111812

Авторы

Космачев Алексей Алексеевич — аспирант, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация; старший преподаватель, Национальный исследовательский университет «Высшая школа экономики», Москва, 123458, Российская Федерация, https://orcid.org/0009-0001-7282-4784, a.kosmachev@hse.ru

Котенко Игорь Витальевич — доктор технических наук, профессор, главный научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация; профессор практики, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 15925268000, https://orcid.org/0000-0001-6859-7120, ivkote@comsec.spb.ru

Authors

Aleksei A. Kosmachev — PhD Student, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation; Senior Lecturer, National Research University Higher School of Economics, Moscow, 123458, Russian Federation, https://orcid.org/0009-0001-7282-4784, a.kosmachev@hse.ru

Igor V. Kotenko — D.Sc., Professor, Chief Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Saint Petersburg, 199178, Russian Federation; Professor of Practice, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 15925268000, https://orcid.org/0000-0001-6859-7120, ivkote@comsec.spb.ru

Статья поступила в редакцию 21.07.2026
Одобрена после рецензирования 16.08.2026
Принята к печати 21.09.2026

Received 21.07.2026
Approved after reviewing 16.08.2026
Accepted 21.09.2026

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»