УДК 004.82 https://doi.org/10.26160/2541-8637-2026-17-82-84

АЛГОРИТМЫ ОБРАБОТКИ СЛАБОСТРУКТУРИРОВАННЫХ ДАННЫХ КИБЕРРАЗВЕДКИ

Вульфин А.М., Кириллова А.Д., Луцкович А.И.
Уфимский университет науки и технологий, Уфа

Ключевые слова: кибеpразведка, большая языковая модель, слабоструктурированные данные, RAG, инцидент, MISP.
Аннотация. В работе предложены алгоритмы обнаружения и анализа слабоструктурированных данных кибеpразведки на основе больших языковых моделей и архитектуры Retrieval-Augmented Generation. Разработана модульная система, включающая агенты сбора информации из OSINT-источников, модуль семантической векторизации на базе Qdrant и модель Saiga-Llama3-8B-Instruct для генерации ответов. Эксперименты на датасете из 10000 документов кибеpразведки показали BERTScore $F1 = 0{,}89$, экспертную оценку 3,98/5 и метрику RAGAS = 0,82 при 20-кратном ускорении обработки по сравнению с ручным анализом.

ALGORITHMS FOR PROCESSING SEMI-STRUCTURED CYBER THREAT INTELLIGENCE DATA

Vulfin A.M., Kirillova A.D., Lutskovich A.I.
Ufa University of Science and Technology, Ufa

Keywords: cyber threat intelligence, large language model, semi-structured data, RAG, incident, MISP.
Abstract. The paper proposes algorithms for detecting and analyzing semi-structured cyber threat intelligence data based on large language models and the Retrieval-Augmented Generation architecture. A modular system was developed that includes agents for collecting information from OSINT sources, a semantic vectorization module based on Qdrant, and the Saiga-Llama3-8B-Instruct model for generating responses. Experiments on a dataset of 10000 cyber intelligence documents showed a BERTScore $F1 = 0.89$, an expert score of 3.98/5, and a RAGAS metric of 0.82 with a 20-fold acceleration in processing compared to manual analysis.

Современный ландшафт угроз характеризуется экспоненциальным ростом объемов данных кибеpразведки (cyber threat intelligence, CTI). При этом значительная часть полезной информации (до 80%) поступает из слабоструктурированных источников: форумов, социальных сетей, Telegram-каналов, отчетов об инцидентах и бюллетеней. Существующие TI-платформы (MISP, OpenCTI, коммерческие решения) эффективно работают со структурированными данными, но обрабатывают менее 30% доступного слабоструктурированного контента [1]. Ручной анализ одного документа занимает в среднем 6 минут, что делает невозможной обработку больших потоков данных в реальном времени [2].

Применение больших языковых моделей (LLM) [3] предоставляет новые возможности для автоматизации процесса анализа слабоструктурированных данных. Однако прямое использование LLM сопряжено с риском «галлюцинаций», что недопустимо в задачах кибербезопасности. Кроме того, модели необходимо снабжать актуальной информацией из внутренних и внешних источников, сохраняя при этом конфиденциальность корпоративных данных.

Для преодоления этих ограничений перспективной является технология Retrieval-Augmented Generation (RAG) [4], которая комбинирует семантический поиск по векторной базе данных с генерацией ответа LLM. Такой подход обеспечивает фактическую обоснованность ответов и позволяет интегрировать разнородные источники информации.

Предлагаемая система анализа данных CTI включает в себя LLM с механизмом RAG как часть системы управления данными CTI. Целью ее создания является повышение эффективности анализа слабоструктурированных данных из различных источников с использованием развернутой платформы MISP.

Система построена на микросервисной архитектуре с использованием агентного подхода к сбору данных. Выделены следующие компоненты:

1. **Агенты сбора данных** реализованы на Python с использованием Beautiful Soup, Scrapy и Bot API. Каждый агент ориентирован на определенный тип источника (веб-сайты, API, RSS, Telegram) и работает асинхронно.

2. **Модуль предварительной обработки** нормализует текст (очистка от HTML, детекция языка), выполняет сегментацию на фрагменты размером 700 токенов с перекрытием 100 токенов и извлекает именованные сущности (IoC, домены, IP-адреса, хеши) с помощью регулярных выражений и моделей spaCy.

3. **Векторная база данных Qdrant**: для семантического индексирования используются эмбеддинги BGE-small-en-v1.5. Поиск выполняется по методу HNSW с квантизацией для ускорения.

4. **Модуль генерации ответов (RAG)**: базовая LLM – Saiga-Llama3-8B-Instruct, адаптированная для русскоязычных задач. При запросе система извлекает из векторной БД наиболее релевантные фрагменты и передает их в модель вместе с промптом. Для инференса применяется vLLM с тензорным параллелизмом.

5. **Интеграционный слой** обеспечивает совместимость с форматами STIX/MISP, а также API для взаимодействия с внешними системами (SIEM, TI-платформы).

Система развернута на сервере с 4×Tesla V100, 128 ГБ видеопамяти, 256 ГБ RAM. Оркестрация контейнеров – Kubernetes, мониторинг – Prometheus/Grafana.

Система прошла оценку (табл. 1) на специально созданном наборе данных. В него вошли 10 тыс. документов, собранных за 2022-2023 годы из четырех категорий источников: бюллетени FinCERT (2500), отчеты НКЦКИ (2000), Telegram-каналы (3000) и специализированные форумы (2500). Для каждого документа экспертами (8 специалистов с опытом 5-15 лет) подготовлено по 5 вопросов и эталонных ответов.

Табл. 1. Метрики качества по категориям документов

Категория	BERT F1	RAGAS	Экспертная оценка	Время, с	Индикатор компрометации, %
FinCERT	0,91	0,86	4,2	15,2	93
НКЦКИ	0,89	0,84	4,0	16,4	91
Telegram	0,87	0,81	3,8	18,3	87
Форумы	0,85	0,78	3,7	19,0	84
Общее	**0,89**	**0,82**	**3,98**	**17,1**	**89**

Система обрабатывает документ в среднем за 17,1 секунды, что в 20 раз быстрее ручного анализа (6,2 мин). Сравнение с альтернативными методами (табл. 2) показывает преимущество предложенного подхода.

Табл. 2. Сравнение с альтернативными методами

Метод	Время на док.	Стоимость/1K, $	Точность, %
Ручной анализ	6,2 мин	500	95
Keyword Extract	2,3 мин	50	65
GPT-3.5 без RAG	45 с	2.5	72
Предложенная система	17,1 с	0.12	89

Масштабируемость: при увеличении числа GPU с 1 до 4 производительность растет с 210 до 690 документов в час (эффективность 85%). Анализ затрат на основе цен AWS показывает операционные расходы $1200 в месяц при обработке 1 млн документов, что в 8-40 раз ниже коммерческих аналогов. Экономическая эффективность: ROI 340-738%, срок окупаемости 8-12 месяцев.

Разработанные алгоритмы и архитектура системы позволяют автоматизировать до 95% процессов обработки слабоструктурированных данных threat intelligence, сохраняя высокое качество анализа (BERTScore $F1 = 0,89$, экспертная оценка 3,98/5). Достигнуто 20-кратное ускорение обработки, обеспечена совместимость со стандартами STIX/MISP.

Список литературы

1. ENISA. Threat Intelligence Platforms: Technical Guidelines 2023 [Электронный ресурс]. – Режим доступа: https://www.enisa.europa.eu.
2. ISC2. Cybersecurity Workforce Study 2024 [Электронный ресурс]. – Режим доступа: https://www.isc2.org.
3. Gholami Y. Large Language Models (LLMs) for Cybersecurity: A Systematic Review // World Journal of Advanced Engineering Technology and Sciences. 2024, vol. 13, pp. 057-069.
4. Singh J.P., Agrawal S. Automating Threat Intelligence Analysis with Retrieval Augmented Generation (RAG) for Enhanced Cybersecurity Posture // International Journal of Science and Research. 2024, vol. 13(5), pp. 251-255.

Сведения об авторах:

Вульфин Алексей Михайлович – д.т.н., профессор кафедры вычислительной техники и защиты информации (ВТиЗИ);

Кириллова Анастасия Дмитриевна – к.т.н., доцент кафедры ВТиЗИ;

Луцкович Андрей Иванович – старший преподаватель кафедры ВТиЗИ.