Научная статья
Original article
УДК 004.08

ТЕОРЕТИКО-МЕТОДОЛОГИЧЕСКИЕ ПОДХОДЫ К РАЗРАБОТКЕ И СОВЕРШЕНСТВОВАНИЮ МЕТОДОВ АНАЛИЗА ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ С ЦЕЛЬЮ ОПРЕДЕЛЕНИЯ УГРОЗ ИТ-БЕЗОПАСНОСТИ

THEORETICAL AND METHODOLOGICAL APPROACHES TO THE DEVELOPMENT AND IMPROVEMENT OF METHODS FOR ANALYZING INFORMATION SOURCES IN ORDER TO IDENTIFY THREATS TO IT SECURITY

Тимофеев Роман Владимирович, Магистрант 2-го года обучения, Московский государственный университет связи и информатики

Timofeev Roman Vladimirovich, vijay-453@mail.ru

Аннотация. В публикации обоснована необходимость и возможность проведения интеллектуального анализа источников текстовых данных с целью определения угроз безопасности в сфере информационных технологий. Обоснован перспективный характер и представлены рекомендации по разработке программ по идентификации и классификации угроз ИТ-безопасности на основе методики обработки естественного языка сообщений новостных лент.

Annotation. The publication substantiates the necessity and possibility of conducting an intelligent analysis of text data sources in order to identify security

threats in the field of information technology. The prospective nature is substantiated and recommendations for the development of programs for the identification and classification of IT security threats based on the natural language processing methodology of news feeds are presented.

Ключевые слова: Natural Language Processing (NLP), обработка естественного языка, анализ естественноязыковых источников, ИТ-безопасность, определение информационных угроз, искусственный интеллект

Keywords: Natural Language Processing (NLP), natural language processing, analysis of natural language sources, IT security, identification of information threats, artificial intelligence

Применение современных цифровых технологий, включая искусственный интеллект (ИИ), призвано активно содействовать целям и задачам обеспечения информационной безопасности объектов.

Риск-менеджмент в сфере информационных технологий (ИТ), как и во многих других сферах, в настоящее время может быть усовершенствован за счет вовлечения в процедуру анализа и оценки рисков источников неформализованного характера, таких, как публикуемые онлайн потоки новостей. На основании анализа содержания таких потоков, представляется возможным комплексно идентифицировать состояние безопасности в ИТ-секторе, а также своевременно выявить новые угрозы, для целей разработки мер по их предотвращению [1].

Идея управления безопасностью и рисками в определенных сферах жизнедеятельности на основе анализа содержания новостных потоков не нова, разработки подобного рода, помимо ИТ-сферы, также востребованы, в частности, в сферах управления национальной безопасностью, в финансовом секторе и др. [2]. Между тем, именно в ИТ-секторе возможно обеспечить двойное применение подобного анализа: помимо идентификации общего состояния безопасности, за счет применения специфического инструментария, возникает возможность также распознать сведения о концептуально новых

ИТ-угрозах для упомянутых выше целей реагирования на них. Для этого представляется необходимым:

- во-первых, систематизировать и классифицировать существующие угрозы и на этой основе сформировать и управлять периодически обновляемыми знаниевыми базами;

- во-вторых, применить к анализу новостных потоков релевантную методику, способную с высокой достоверностью и точностью обнаруживать непосредственно сведения об ИТ-угрозах, отграничивая их от иных сведений, а также, сопоставляя с данными существующих учетных регистров, идентифицировать такие угрозы в качестве новых.

С учетом постановки задачи, ее решение представляется возможным осуществить на основе применения такой методики, как обработка естественного языка (англ. Natural Language Processing (NLP)) [3].

NLP представляет собой современное направление математической лингвистики, а также систем искусственного интеллекта. Методология основана на сочетании методов компьютерного анализа массивов, логически связанных текстовых данных и синтеза естественных языков. При обеспечении понимания искусственным интеллектом естественного языка, достигается полноценное решение задачи функционирования и развития ИИ, поскольку главный смысл таких технологий выступает в полноценном понимании и распознавании объектов, включая элементы живого языка.

Разработки технологий распознавания и обработки живых языков проводятся длительное время. В качестве длительно и устойчиво доказавших свою утилитарность, могут быть названы такие технологии-предшественники NLP, как интерактивное голосовое меню (IVR).

В настоящее время, разработки систем NLP продвинулись от индивидуальных к кастомизированным разработкам, к конструкторам и платформам для создания NLP-программ [4]. В качестве широко используемой в практической деятельности, может быть названа NLTK (Natural Language Toolkit) платформа для разработки программ NLP на основе

Python, распространяемая как открытый источник и характеризующаяся наличием утилитарных и дружественных интерфейсов для значительно числа языков, библиотек, используемых для различных целей. Из аналогичных актуальных платформ для разработки NLP-программ на Python может быть упомянут Stanza – инструмент для обработки естественных языков с открытым исходным кодом, поддерживающий 66 естественных языков [5].

С учетом состояния развития технологий анализа естественноязыковых источников, для решения поставленной задачи содействия обеспечению ИТ-безопасности на основе мониторинга корпуса новостных текстов, может быть рекомендован нижеследующий концептуальный подход к разработке аналитической методологии. С применением NLP-программы, предназначенной для отслеживания определенных потоков текстовой информации, предпочтительно – новостных лент СМИ (как источника концентрированной актуальной и сравнительно релевантной информации), на основе использования приемов токенизации по словам и предложениям, лемматизации и стемминга текста, регулярных выражений и «стоп»-слов, «мешка слов» и частот слов и корпуса слов, представляется возможным с высокой точностью идентифицировать описание новых ИТ-угроз (вредоносного ПО, уязвимостей и т.п.) и составлять на каждую угрозу карточку с ее описанием. Полученные результаты выступят основой для разработки частных и общих мер реагирования на соответствующие угрозы.

Таким образом, на основе методики обработки естественного языка сообщений новостных лент, могут быть разработаны перспективные программы по идентификации и классификации угроз ИТ-безопасности.

Список литературы

1. Rodríguez A., Ortega F., Concepción R. An intuitionistic method for the selection of a risk management approach to information technology projects //Information Sciences. – 2017. – Vol. 375. – Pp. 202-218.
2. Близнюк Б. О. и др. Современные методы обработки естественного языка //Вестник Харьковского национального университета имени В.Н.

Каразина. Серия: Математическое моделирование. Информационные технологии. Автоматизированные системы управления. – 2017. – №. 36. – С. 14-26.

3. Chowdhary K. R. Natural language processing //Fundamentals of Artificial Intelligence. – New Delhi: Springer, 2020. – Pp. 603-649.
4. Colton D. Text Classification Using Python //Text Mining and Visualization: Case Studies Using Open-Source Tools. – 2016. – Vol. 40. – P. 199.
5. Qi P. et al. Stanza: A python natural language processing toolkit for many human languages //arXiv preprint arXiv:2003.07082. – 2020.

List of literature

1. Rodríguez A., Ortega F., Concepción R. An intuitionistic method for the selection of a risk management approach to information technology projects //Information Sciences. - 2017. - Vol. 375. - Pp. 202-218.
2. Bliznyuk B. O. et al. Modern methods of natural language processing //Bulletin of V.N. Karazin Kharkiv National University. Series: Mathematical modeling. Information technology. Automated control systems. - 2017. - No. 36. - pp. 14-26.
3. Chowdhary K. R. Natural language processing //Fundamentals of Artificial Intelligence. – New Delhi: Springer, 2020. – Pp. 603-649.
4. Colton D. Text Classification Using Python //Text Mining and Visualization: Case Studies Using Open-Source Tools. – 2016. – Vol. 40. – P. 199.
5. Qi P. et al. Stanza: A python natural language processing toolkit for any human languages //archive preprint arXiv:2003.07082. - 2020.

© Тимофеев Роман Владимирович Научно-образовательный журнал для студентов и преподавателей «StudNet» №3/2022.

Для цитирования: Тимофеев Р.В. Теоретико-методологические подходы к разработке и совершенствованию методов анализа информационных источников с целью определения угроз ИТ-безопасности // Научно-образовательный журнал для студентов и преподавателей «StudNet» №3/2022