УДК 004.056, 004.934, 004.912

ПОВЫШЕНИЕ ТОЧНОСТИ КЛАССИФИКАЦИИ УЯЗВИМОСТЕЙ И ОПРЕДЕЛЕНИЕ УЯЗВИМОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ НЕСТРУКТУРИРОВАННОГО ОПИСАНИЯ

Нкодиа Д.-К. (Университет ИТМО) **Менщиков А. А.** (Университет ИТМО)
Научный руководитель – к.т.н., доцент Менщиков А.А.
(Университет ИТМО)

В работе протестированы методы классификации текста. Предложен алгоритм, повышающий точность классификации уязвимостей по их неструктурированному текстовому описанию. Рассмотрены подходы по определению названий уязвимого программного обеспечения.

Для поддержания информационной безопасности необходимо регулярно проводить ревизию используемых компонентов системы и соответствующих им уязвимостей. Учет недостатков программного обеспечения (далее – ПО) является трудоемкой задачей из-за постоянного появления новой неструктурированной информации об обнаруженных уязвимостях. Для ускорения процесса анализа релевантной информации была поставлена задача разработать алгоритм обработки описания уязвимостей. Под обработкой текстового описания подразумевается выявление названия и версии уязвимого ПО, а также определение класса уязвимости и уровня опасности. Эффективность алгоритма предлагается оценивать в соответствии с точностью классификации. В связи с этим подзадачей исследования является выбор метода, с помощью которого достигается наибольшая точность классификации.

С целью определить наиболее точный метод классификации были протестированы два алгоритма машинного обучения (метод опорных векторов, SVM, и случайный лес, Random Forest) и метод на основе нейронной сети (рекуррентная нейронная сеть с долгой краткосрочной памятью, LSTM RNN). Алгоритмы SVM и LSTM RNN были выбраны для сравнения, поскольку они продемонстрировали высокие показатели точности в изученных исследованиях, так же посвященных анализу информации, связанной с информационной безопасностью. Кроме того, проводилась дополнительная оптимизации параметров классификатора LSTM RNN для повышения точности результатов. Тестирование алгоритма Случайный лес было проведено в силу того, что данный метод является наиболее часто применимым для задач мультиклассовой классификации, и при этом требующий меньше ресурсов чем нейронные сети.

Также в ходе разработки алгоритма были сравнительно проанализированы подходы по определению названий уязвимого ПО, основанные на применении словаря и методах искусственного интеллекта. Более универсальными, но и ресурсоемкими подходами по результатам сравнения являются методы извлечения именованных сущностей (NER).

Разработанной алгоритм основывается на наиболее эффективных методах классификации и может определять название уязвимого ПО, позволяя ускорить процесс обработки информации о новых уязвимостях.

Нкодиа Д.-К. (автор) _____

Менщиков А.А. (научный руководитель) _____