УДК 004.896

ПОДХОД К СБОРУ И ОБРАБОТКЕ ИНФОРМАЦИИ ДЛЯ ОБНАРУЖЕНИЯ НЕГАТИВНЫХ КОММЕНТАРИЕВ И СООБЩЕНИЙ ИЗ РАЗЛИЧНЫХ ИНТЕРНЕТ-ИСТОЧНИКОВ

Рябченко В.А. (ВКА)**, Петрухина А.А.** (ВКА)**, Дудкин А.С.** (ВКА)
Научный руководитель – кандидат технических наук, доцент Дудкин А.С.
(Военно-космическая академия имени А.Ф.Можайского)

Введение. Современные онлайн-платформы играют важную роль в распространении информации, формировании общественного мнения и взаимодействии между людьми по всему миру. Социальные сети, мессенджеры и видеохостинги привлекают миллионы пользователей благодаря возможности обмениваться новостями, публиковать контент и участвовать в дискуссиях на актуальные темы. Однако, несмотря на их значимость, данные платформы сталкиваются с рядом вызовов, включая высокую конкуренцию, стремительный рост аудитории и необходимость соблюдения норм конфиденциальности.

В условиях усиливающихся ограничений на доступ к данным, платформы вводят механизмы защиты, такие как обязательная регистрация, ограничение количества запросов и использование капч. Эти меры затрудняют автоматизированный сбор данных, необходимый для анализа информационных потоков, мониторинга общественного мнения и выявления негативного контента. Тем не менее, эффективное использование современных технологий позволяет преодолеть эти трудности и получить доступ к открытой информации для её последующей обработки.

Основная часть. Сбор данных с онлайн-платформ требует преодоления различных технических и организационных ограничений. Одним из ключевых инструментов для автоматизации этого процесса является Selenium WebDriver [4]. Этот инструмент позволяет имитировать действия пользователя в браузере, включая ввод текста, клики, прокрутку страниц и взаимодействие с элементами интерфейса. Благодаря этому возможно автоматизировать процесс получения данных, таких как текстовый контент, изображения, видеоматериалы и комментарии.

Selenium особенно полезен в случаях, когда платформы не предоставляют открытых API для доступа к данным или ограничивают их функционал [4]. Например, инструмент может использоваться для обхода интерфейсов, требующих регистрации, или для выполнения сложных действий, таких как фильтрация контента по заданным параметрам. При этом Selenium работает с большинством современных браузеров, что делает его универсальным решением для автоматизации задач.

Однако одной из сложностей при работе с социальными сетями и мессенджерами является необходимость повторной авторизации при каждом запуске программы, что связано с использованием платформами механизмов безопасности, таких как cookies и local storage, где хранятся данные о сессии пользователя. Чтобы избежать повторного ввода учетных данных, эти данные можно сохранить с помощью модуля pickle в Python. Pickle позволяет сериализовать объекты Python, такие как cookies, и восстанавливать их для повторного использования.

Комбинация Selenium и pickle открывает широкие возможности для автоматизации мониторинга данных. Например, можно настроить регулярный сбор информации с заданных страниц, анализировать динамику публикаций и выявлять ключевые паттерны в комментариях. Это особенно актуально для анализа негативного контента, направленного против Российской Федерации, или мониторинга информационных потоков в реальном времени.

Кроме того, для повышения эффективности сбора данных можно использовать дополнительные инструменты, такие как библиотеки BeautifulSoup или Scrapy, которые

позволяют обрабатывать HTML-код и извлекать нужные элементы [5]. В сочетании с Selenium эти технологии дают возможность работать даже с динамически загружаемым контентом, что значительно расширяет спектр задач, которые можно решать.

В анализе данных также можно использовать современные технологии обработки естественного языка, такие как модели трансформеров [3], и методы машинного обучения [6], которые позволяют выявлять тональность сообщений, анализировать лексические конструкции и автоматически классифицировать контент.

Выводы. Применение современных инструментов, таких как Selenium WebDriver и модуль pickle, позволяет автоматизировать сбор данных с онлайн-платформ, обходя существующие ограничения. Эти технологии обеспечивают доступ к необходимой информации, создавая возможности для анализа контента, мониторинга информационных потоков и выявления деструктивно направленной риторики в сторону Российской Федерации.

Список использованных источников:

1. Харитонов А.В. Современные методы анализа интернет-контента: подходы и технологии. – Москва: Наука, 2023.
2. Соловьев А.А., Иванова Е.П. Машинное обучение в анализе текстовых данных. – Санкт-Петербург: Питер, 2022.
3. Vaswani, A., Attention Is All You Need. 2017. URL: https://arxiv.org/abs/1706.03762 (дата обращения: 2.10.2024).
4. Selenium WebDriver Documentation. URL: https://www.selenium.dev/documentation/ (дата обращения: 3.11.2024).
5. Richardson, L. Beautiful Soup Documentation. URL: https://www.crummy.com/software/BeautifulSoup/ (дата обращения: 5.11.2024).
6. Bishop, C. M. Pattern Recognition and Machine Learning. – Springer, 2006.

Рябченко В.А. (автор) Подпись

Петрухина А.А. (автор) Подпись

Дудкин А.С. (автор) Подпись