УДК 004.622

ПРОЗРАЧНОСТЬ ОБУЧАЮЩИХ ДАННЫХ КАК ФАКТОР ИНТЕРПРЕТИРУЕМОСТИ AI-ТЕХНОЛОГИЙ

Манохин К. А.[1], Барахсин Г. М.[1]
Научный руководитель – Шалыгин В. А.[1]
[1]Университет ИТМО
kirillmanokhin@gmail.ru

Работа выполнена в рамках темы НИР №625111 «Алгоритмическая прозрачность и предиктивные модели развития AI-технологий: от базы знаний к навигатору решений».

Введение

В настоящее время технологии искусственного интеллекта (ИИ) активно применяются в качестве систем аналитики, поддержки принятия управленческих решений, а также в социально-экономических исследованиях. В свою очередь, данные, оставляемые пользователями, в частности отзывы и цифровые следы, служат источником информации о восприятии продуктов и услуг. Современные исследования демонстрируют, что обеспечение прозрачности обучающих данных способствует повышению интерпретируемости результатов, снижению галлюцинаций моделей ИИ и, как следствие, укреплению доверия к алгоритмическим решениям. [1, 2]. Однако во многих современных системах ИИ уровень прозрачности наборов данных остается низким, что снижает возможность проверки и воспроизведения результатов [1]. В то же время задачи документирования, очистки и контроля качества обучающих данных рассматриваются как необходимые условия для повышения прозрачности исследований в области ИИ [3].

Основная часть

Недостаточная прозрачность обучающих данных прежде всего проявляется в отсутствии фиксированной информации об их происхождении и предобработке, что не позволяет оценить, являются ли данные реальными или синтетическими [1, 2, 4]. В качестве решения этой проблемы авторами предлагается методология формирования корпуса обучающих данных, которая обеспечивает прозрачность их происхождения. Методология предусматривает:

1. Фиксацию источника данных по его адресу URL и временной метке — фиксируется дата и время, когда отзыв был получен системой.
2. Формирование контрольных хеш-сумм в формате SHA-256 для обеспечения уникальности каждого отзыва при проведении дедупликации, а также для возможности его однозначной идентификации.
3. Возможность адаптировать программное обеспечение под различные типы собираемых данных из открытых веб-платформ с динамическим обновлением контента, что увеличивает объем получаемых обучающих данных.

Обучающие данные сохраняются в формате JSONL вместе с их метаданными, позволяющими однозначно идентифицировать происхождение всего корпуса. На основе предложенного подхода реализован и протестирован программный парсер для сбора пользовательских отзывов из геоинформационных сервисов. Тестирование показало устойчивую работоспособность системы и возможность повторного воспроизведения содержимого корпуса данных при сохранении целостности текстовых отзывов.

Выводы

Разработана и протестирована методология парсинга пользовательских отзывов из геоинформационных сервисов, обеспечивающая фиксацию происхождения

обучающих данных. Реализованный программный парсер продемонстрировал устойчивую работоспособность и воспроизводимость формируемого корпуса данных. Полученные результаты подтверждают возможность повышения прозрачности обучающих наборов данных на этапе их формирования.

Литература

1. AI data transparency: an exploration through the lens of AI incidents / S. Worth [et al.] // arXiv / Computers and Society. — 2024. — arXiv:2409.03307
2. Data Quality in the Age of AI: A Review of Governance, Ethics, and the FAIR Principles / M. Guillen-Aguinaga [et al.] // Data. — 2025. — Vol. 10. — No. 12. — 201.
3. S. Zulaikha, I. R. Dewi, M. Kurniawati. Unveiling the intellectual landscape of artificial intelligence and consumer behavior // Discover Artificial Intelligence. — 2026. — Vol. 6. — 2.
4. A. Adanyin. Ethical AI in Retail: Consumer Privacy and Fairness // arXiv / Computers and Society. — 2024. — arXiv:2410.15369