**УДК 004.89**

# ПРИМЕНЕНИЕ ТЕХНИК АНОНИМИЗАЦИИ ДЛЯ СНИЖЕНИЯ РИСКА УТЕЧЕК В ОБУЧЕННЫХ МОДЕЛЯХ

Гладких Д.С., младший научный сотрудник
Военная академия связи имени Маршала Советского Союза
С. М. Буденного, г. Санкт-Петербург

Анонимизация данных, применяемая в процессе подготовки обучающих выборок для моделей машинного обучения, представляет собой один из ключевых методов снижения рисков утечки информации, особенно в задачах, связанных с персональными, медицинскими, финансовыми или поведенческими данными. Несмотря на то, что анонимизация часто воспринимается как предварительный этап обработки данных, её роль выходит далеко за рамки начальной фильтрации: корректное и устойчивое применение техник анонимизации оказывает прямое влияние на свойства модели – её обобщающую способность, устойчивость к атакам и соответствие требованиям доверия и нормативной совместимости [1].

Актуальность этой темы возрастает в условиях использования больших предобученных моделей, в которых сложно отследить, какие фрагменты данных были запомнены или могут быть восстановлены при соответствующих атаках, особенно через API или при открытом доступе к модели.

Классические методы анонимизации, такие как удаление идентификаторов (имён, номеров, адресов), замена редких категорий, обобщение по уровням таксономий (например, замена точного возраста на возрастную группу), используются достаточно давно и активно интегрированы в инструменты предобработки данных. Однако в контексте обучения моделей эти методы оказываются зачастую недостаточными: нейросети, особенно глубокие архитектуры, способны извлекать информацию из косвенных признаков, прокси-атрибутов и статистических ассоциаций, которые сохраняются даже после поверхностной деидентификации. Таким образом, простая анонимизация не гарантирует защиту от восстановления чувствительных характеристик [2].

В ответ на это были разработаны более формальные подходы, среди которых центральное место занимает k-анонимность – требование, чтобы каждый индивидуум в датасете не отличался от не менее чем $k-1$ других по комбинации квазиидентификаторов. В то время как эта модель полезна в табличных данных, она плохо масштабируется к непрерывным и высокоразмерным пространствам, характерным для компьютерного зрения или обработки текста.

Более того, соблюдение k-анонимности может нарушаться в результате обучения модели, если веса модели «запоминают» характерные паттерны отдельных индивидуумов. Для борьбы с этим применяются техники локального сглаживания, синтетической генерации данных и кластерной маскировки, которые стремятся обеспечить схожесть представлений на уровне признаков, а не только на уровне исходных значений [3].

Другим направлением является использование d-анонимизации и l-диверсификации, где гарантируется не только неотличимость, но и разнообразие значений чувствительных атрибутов в каждом кластере анонимизированных записей. Это уменьшает вероятность успешного восстановления конкретных значений при атаке на обученную модель. Однако такие методы требуют строгого контроля распределения данных и могут вызывать значительное искажение информации, снижая точность модели.

Современные методы анонимизации всё чаще опираются на генеративные подходы, например, с использованием вариационных автокодировщиков (VAE), GAN или диффузионных моделей, которые обучаются создавать новые примеры данных, статистически близкие к исходным, но не соответствующие реальным индивидуумам. Такой синтетический датасет может использоваться для обучения моделей, снижая вероятность запоминания реальных чувствительных записей. Эти методы обладают высокой гибкостью, позволяют управлять уровнем разнообразия, но также требуют оценки уровня «остаточной идентифицируемости» – то есть возможности отнести сгенерированные примеры к конкретным субъектам или группам, особенно если используются в задачах с ограниченной структурой классов [4].

В задачах обработки естественного языка (NLP) применяются специальные методы маскировки имён собственных, дат, организаций и других сущностей, используя регулярные выражения, модели named entity recognition (NER), или контекстные эмбеддинги. Однако исследования показывают, что даже после замены именованных сущностей LLM могут сохранять стилистические и тематические особенности текстов, позволяющие атакующему восстановить автора или контекст при наличии вспомогательной информации. Визуальные модели, в свою очередь, требуют удаления или искажения биометрических признаков – лиц, татуировок, уникальных предметов одежды, что может достигаться как традиционными методами размывания и замены, так и современными методами "inpainting" или генеративного редактирования изображений.

Отдельное внимание уделяется анонимизации в контексте федеративного обучения, где данные остаются на устройствах пользователей, а модель обучается через обмен градиентами или обновлениями весов. Здесь требуется анонимизация не только самих данных, но и коммуникационных каналов, чтобы предотвратить корреляцию обновлений с конкретными клиентами.

Используются методы shuffle-механизмов, рандомизации маршрутов обмена, а также добавление шума к обновлениям (см. например, secure aggregation и local differential privacy) [5].

Критически важным аспектом остаётся оценка качества анонимизации. Для этого применяются метрики, измеряющие баланс между сохранением информативности данных и уровнем приватности: это могут быть показатели mutual information, re-identification risk, delta-disclosure risk, а также качество предсказаний модели после анонимизации. В некоторых случаях проводится автоматический аудит: проверка, насколько хорошо обученная модель может предсказывать или реконструировать анонимизированные атрибуты. Если предсказательная сила сохраняется, значит информация в латентном виде всё ещё кодируется [6].

Таким образом, применение техник анонимизации в контексте доверенного машинного обучения требует не просто удаления идентификаторов, а глубокого анализа структуры признаков, латентных зависимостей, архитектур модели и сценариев использования. Эффективная анонимизация – это многоуровневая стратегия, сочетающая формальные методы, эмпирическую проверку и генеративные подходы, обеспечивающая защиту данных пользователей без существенного ущерба для производительности ИИ-системы.

В условиях ужесточения нормативных требований и повышения социальной значимости ИИ-сервисов такие подходы становятся не просто желательными, а обязательными элементами построения ответственного и доверенного искусственного интеллекта [7].

**Список литературы:**

1. Yousefpour, A. Opacus: User-Friendly Differential Privacy Library in PyTorch / A. Yousefpour, I. Shilov, A. Sablayrolles [et al.] // arXiv preprint. – 2021. – arXiv:2109.12298.

2. Holohan, N. Diffprivlib: The IBM Differential Privacy Library / N. Holohan, S. Braghin, P. Mac Aonghusa, K. Levacher // arXiv preprint. – 2019. – arXiv:1907.02444.

3. TensorFlow. TensorFlow Privacy: Responsible AI Toolkit [Электронный ресурс]. – URL: https://www.tensorflow.org/responsible_ai/privacy/guide (дата обращения: 09.06.2025).

4. Meta AI. Opacus: Training PyTorch Models with Differential Privacy [Электронный ресурс]. – URL: https://github.com/meta-pytorch/opacus (дата обращения: 09.06.2025).

5. Google. TensorFlow Federated: Machine Learning on Decentralized Data [Электронный ресурс]. – URL: https://www.tensorflow.org/federated (дата обращения: 09.06.2025).

6. Sweeney, L. k-Anonymity: A Model for Protecting Privacy / L. Sweeney // International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. – 2002. – Vol. 10, № 5. – P. 557–570. – DOI 10.1142/S0218488502001648.

7. Machanavajjhala, A. l-Diversity: Privacy Beyond k-Anonymity / A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam // ACM Transactions on Knowledge Discovery from Data. – 2007. – Vol. 1, № 1. – Article 3. – DOI 10.1145/1217299.1217302.