УДК *004.8:316.472.4*

Ямкин Сергей Геннадьевич,
ООО «ВБ ТЕХ», отдел DevOps Monitoring, DevOps-инженер,
г. Москва,

ИНТЕЛЛЕКТУАЛЬНАЯ МОДЕРАЦИЯ КОНТЕНТА В СОЦИАЛЬНОЙ СЕТИ: АРХИТЕКТУРНЫЕ РЕШЕНИЯ И АЛГОРИТМИЧЕСКИЕ ПОДХОДЫ

Аннотация. *В статье рассматриваются подходы к построению системы интеллектуальной модерации контента в социальной сети. Анализируются архитектурные принципы интеграции модулей искусственного интеллекта в инфраструктуру платформы, а также алгоритмические методы обработки текстовой и визуальной информации. Особое внимание уделено вопросам масштабируемости, скорости обработки пользовательского контента и снижению нагрузки на ручную модерацию. На основе анализа научных источников сформулированы требования к построению архитектуры системы автоматической фильтрации контента, включая современные подходы к мультимодальному анализу и организации обратной связи для непрерывного обучения моделей.*

Ключевые слова: *социальная сеть, модерация контента, искусственный интеллект, машинное обучение, обработка естественного языка, компьютерное зрение, архитектура информационных систем, микросервисы, масштабируемость.*

Sergey G. Yamkin,
VB TECH LLC, DevOps Monitoring Department, DevOps Engineer,

Moscow,

e-mail: rdzvl@yandex.ru

INTELLIGENT CONTENT MODERATION IN SOCIAL NETWORK: ARCHITECTURAL SOLUTIONS AND ALGORITHMIC APPROACHES

Abstract. *The article discusses approaches to building an intelligent content moderation system in a social network. The architectural principles of integrating artificial intelligence modules into the platform infrastructure, as well as algorithmic methods for processing textual and visual information are analyzed. Special attention is paid to scalability, speed of user content processing and reducing the load on manual moderation. Based on the analysis of scientific sources, requirements for building the architecture of an automated content filtering system are formulated, including modern approaches to multimodal analysis and organization of feedback for continuous learning of models.*

Keywords: *social network, content moderation, artificial intelligence, machine learning, natural language processing, computer vision, information system architecture, microservices, scalability.*

Введение

Тема исследования. Интеллектуальная модерация контента в архитектуре социальной сети.

Цель исследования. Проанализировать архитектурные и алгоритмические подходы к созданию системы автоматической модерации пользовательского контента на основе методов искусственного интеллекта.

Проблема исследования. Рост объёма пользовательского контента в социальных сетях делает невозможной его эффективную проверку исключительно силами модераторов-людей. Это приводит к распространению нежелательных материалов, спама, оскорбительного и запрещённого

контента. Внедрение интеллектуальных методов модерации требует разработки архитектуры, обеспечивающей быструю обработку данных, высокую точность классификации и устойчивость к нагрузкам.

Метод исследования. Метод исследования основан на анализе научной и учебной литературы, посвящённой обработке естественного языка, компьютерному зрению, машинному обучению и архитектуре распределённых информационных систем. Также рассматриваются практические кейсы внедрения подобных систем в крупных социальных платформах.

Основная часть

1. Современные вызовы и необходимость интеллектуальной модерации

Современные социальные сети ежедневно обрабатывают миллионы публикаций, комментариев, изображений и видеоматериалов. Объём информации настолько велик, что ручная модерация становится вспомогательным, а не основным инструментом контроля контента [1, с. 45]. В этих условиях особую роль приобретает интеллектуальная модерация — автоматизированная система анализа данных, основанная на алгоритмах искусственного интеллекта [2].

Развитие систем ИИ-модерации проходит в условиях постоянной «гонки вооружений» с пользователями, пытающимися обойти фильтры. Это требует создания не просто статических правил, а адаптивных, самообучающихся систем, способных выявлять новые паттерны нарушений. Ключевыми вызовами при этом становятся не только точность классификации, но и обеспечение бесперебойной работы под высокой нагрузкой, минимизация ложных срабатываний (over-blocking) и соблюдение этических норм при автоматическом принятии решений [3, с. 210]. Особую сложность представляют контекстно-зависимые нарушения, где одно и то же сообщение

может быть как допустимым, так и запрещённым в зависимости от культурного контекста, аудитории и текущих событий [4].

2. Архитектурные принципы построения системы модерации

С архитектурной точки зрения система модерации должна быть интегрирована в общий контур обработки пользовательских данных. Контент проходит через несколько этапов: загрузка, предварительная проверка, интеллектуальный анализ и принятие решения о публикации. Такой конвейер позволяет разделить задачи по уровням сложности и уменьшить задержку при публикации допустимого контента [5, с. 312].

Типичный конвейер обработки включает следующие компоненты:

Приём и буферизация контента: Использование высокопроизводительных брокеров сообщений (например, Apache Kafka, RabbitMQ) для обработки пиковых нагрузок и обеспечения отказоустойчивости. На этом этапе происходит нормализация данных и их распределение по очередям в зависимости от типа контента [6, с. 178].

Статический анализ и фильтрация (Pre-filtering): Быстрая проверка по хэш-суммам (борьба с известными запрещёнными материалами через системы типа PhotoDNA), регулярным выражениям, чёрным и белым спискам. Этот слой отсеивает до 40-60% очевидных нарушений с минимальными вычислительными затратами.

Интеллектуальный анализ: Параллельный или последовательный запуск специализированных моделей для текста, изображений, видео, аудио и метаданных. Архитектура должна обеспечивать изоляцию падений отдельных моделей и graceful degradation при сбоях [7].

Aggregation & Decision Engine: Сервис, агрегирующий оценки от всех моделей (например, с использованием взвешенного голосования, мета-классификатора или экспертной системы правил) и принимающий конечное решение на основе настроенных политик платформы. Здесь же применяются бизнес-логика и контекстные правила.

Действие и логирование: Применение решения (пропуск, блокировка, пометка, отправка на ручную проверку) с обязательным сохранением всего контекста для аудита, анализа эффективности и последующего дообучения моделей.

3. Алгоритмические подходы к анализу текстового контента

Первым уровнем является фильтрация на основе простых правил. Она отсеивает очевидные нарушения, например запрещённые слова или повторяющийся спам. Однако этот подход неэффективен против завуалированных форм нарушений, что требует применения методов машинного обучения [8, с. 567].

Для анализа текстового контента применяются алгоритмы обработки естественного языка [4]. Модели классификации текста позволяют выявлять оскорбления, призывы к насилию, дезинформацию и другие типы нежелательных сообщений. Архитектурно текстовый анализ реализуется как отдельный сервис, принимающий текстовые данные через API и возвращающий вероятность принадлежности сообщения к определённому классу.

Современные подходы к анализу текста эволюционировали от классических методов (например, SVM с TF-IDF, наивный байесовский классификатор) к глубокому обучению [2, с. 450]. На сегодняшний день доминируют трансформерные архитектуры, такие как BERT и его производные (например, RoBERTa, DeBERTa, Multilingual BERT), предобученные на больших корпусах текстов и способные учитывать глубокий контекст [9]. Это критически важно для распознавания сарказма, эвфемизмов, иронии и культурно-зависимых выражений. Для задач модерации создаются специализированные дообученные версии этих моделей на размеченных датасетах, содержащих примеры токсичных высказываний, дезинформации, экстремистского контента и т.д.

Особое значение имеет анализ семантической близости и векторные представления слов (word embeddings), которые позволяют выявлять

синонимичные нарушения и адаптироваться к изменяющейся лексике. Современные системы также начинают учитывать прагматику и дискурс-анализ для понимания намерений пользователя в рамках целой дискуссии, а не отдельного сообщения [3, с. 215].

4. Методы компьютерного зрения и анализ мультимедийного контента

Обработка изображений и видео требует использования методов компьютерного зрения [10, с. 89]. Нейронные сети способны распознавать запрещённые объекты, сцены насилия или откровенный контент. Эти вычислительно сложные операции выполняются на специализированных серверах, что требует выделения отдельного вычислительного контура в архитектуре системы.

Для анализа изображений наиболее эффективными оказались свёрточные нейронные сети (CNN) архитектур ResNet, EfficientNet, Vision Transformers (ViT) [11]. Эти модели обучаются на миллионах размеченных изображений и способны выявлять как явные нарушения (насилие, нагота), так и более сложные концепции (символику экстремистских организаций, контент, связанный с самоповреждением).

Анализ видео представляет особую сложность, так как требует обработки не только пространственных, но и временных признаков. Для этого применяются 3D-свёрточные сети (3D-CNN) или гибридные архитектуры, сочетающие 2D-CNN для анализа кадров и рекуррентные сети (RNN/LSTM) для анализа последовательностей [12]. Также активно используются методы эффективного сэмплирования ключевых кадров для снижения вычислительной нагрузки. Для обнаружения сложных контекстных нарушений (например, буллинга, опасных челленджей, деструктивного поведения) системы начинают применять мультимодальный анализ, объединяя данные из видеоряда, аудиодорожки и субтитров.

Важным направлением является обнаружение манипуляций с контентом (deepfakes, монтаж) с помощью методов обнаружения артефактов генерации и анализа временных несоответствий [13].

5. Анализ метаданных и графовых взаимосвязей

Помимо анализа непосредственно контента, критически важным источником информации являются метаданные и граф социальных взаимодействий [14, с. 332]. Анализ поведения пользователя (частота публикаций, паттерны комментариев, история нарушений, время активности), его связей и участия в сообществах позволяет выявлять скоординированные атаки, ботовые сети и устойчивых нарушителей.

Для этого в архитектуру интегрируются графовые нейронные сети (Graph Neural Networks, GNN), способные эффективно обрабатывать связанные данные и выявлять подозрительные кластеры активности, что невозможно при изолированном анализе отдельного поста [15]. Алгоритмы обнаружения сообществ (community detection) и анализа распространения информации (information diffusion) помогают идентифицировать организованные группы, распространяющие вредоносный контент.

Анализ метаданных включает проверку геолокации, устройств, IP-адресов, шаблонов поведения, что особенно эффективно против спам-атак и создание фейковых аккаунтов. Системы репутационного скоринга, основанные на совокупной истории активности пользователя, позволяют применять персонализированные пороги срабатывания фильтров [16, с. 78].

6. Организация асинхронной обработки и масштабируемость

Важным архитектурным решением является использование асинхронной обработки [5, с. 324]. Контент может быть опубликован с ограничениями (например, только для подписчиков) до завершения полной проверки или отправляться на постмодерацию. Такой подход позволяет сохранять высокую скорость работы платформы при одновременном обеспечении безопасности. Реализуется это через механизмы отложенных задач (delayed jobs) и очередей с приоритетами.

Система модерации должна быть масштабируемой, поскольку нагрузка на неё напрямую зависит от активности пользователей. Микросервисная архитектура позволяет масштабировать модули текстового и визуального

анализа независимо друг от друга [6, с. 201]. Контейнеризация (Docker) и оркестрация (Kubernetes) обеспечивают гибкое управление ресурсами и отказоустойчивость.

Для обработки пиковых нагрузок применяются стратегии:

Динамическое масштабирование (autoscaling) по CPU/памяти или длине очереди;

Географическая дистрибуция обработки данных;

Кэширование результатов проверки похожего контента;

Sampling (выборочная проверка) для низкорисковых категорий контента или пользователей с высокой репутацией.

7. Обратная связь и непрерывное обучение системы

Необходимо учитывать возможность ошибок алгоритмов. Поэтому архитектура включает механизм передачи спорных случаев на ручную проверку. Результаты работы модераторов используются для дообучения моделей, что повышает точность системы со временем [7].

Таким образом, формируется непрерывный цикл улучшения (MLOps Pipeline):

Сбор данных: автоматический сбор спорных кейсов (пограничных решений), новых типов нарушений, жалоб пользователей.

Разметка данных: создание тренировочных выборок силами модераторов с применением активного обучения (active learning) для приоритизации наиболее информативных примеров [17, с. 290].

Дообучение моделей: инкрементальное обучение или полное переобучение моделей на расширенных датасетах с контролем дрейфа данных (data drift).

Валидация и тестирование: A/B-тестирование новых версий моделей на изолированном трафике, оценка метрик качества [16, с. 154].

Деплоймент: постепенный роллаут (canary release, blue-green deployment) новых моделей в продакшен.

Мониторинг: отслеживание метрик в реальном времени, детектирование аномалий и деградации качества.

Для оценки эффективности системы используются не только технические метрики (точность, полнота, F1-мера, ROC-AUC), но и бизнес-показатели: снижение количества жалоб пользователей, время реакции на новые угрозы, процент ложных блокировок, нагрузка на ручных модераторов.

8. Этические аспекты и прозрачность решений

Интеллектуальная модерация контента представляет собой комплекс программных и алгоритмических решений, обеспечивающих автоматический анализ пользовательских данных и поддержание безопасной цифровой среды. Однако автоматическое принятие решений о блокировке контента порождает серьёзные этические вопросы, связанные с цензурой, предвзятостью алгоритмов и правом на апелляцию [18].

Современные системы должны включать механизмы explainable AI (XAI), позволяющие понять, на основе каких признаков было принято решение [19]. Это важно как для внутреннего аудита, так и для предоставления обратной связи пользователям. Архитектура должна предусматривать прозрачный процесс обжалования решений и человеческий надзор за спорными случаями.

Проблема алгоритмической предвзятости (bias) требует особого внимания при обучении моделей и формировании датасетов. Несбалансированные или нерепрезентативные данные могут привести к дискриминации определенных групп пользователей или точек зрения [3, с. 225].

Заключение

Интеллектуальная модерация является неотъемлемой частью современной социальной сети. Её архитектура должна обеспечивать масштабируемость, высокую скорость обработки данных и возможность

постоянного улучшения моделей. Использование методов машинного обучения и компьютерного зрения позволяет существенно снизить нагрузку на ручных модераторов и повысить качество фильтрации контента.

Перспективы развития лежат в области более глубокой мультимодальности (совместный анализ текста, изображения, звука и поведения в единой модели), explainable AI (XAI) для обеспечения прозрачности решений и снижения предвзятости алгоритмов, а также федеративного обучения, позволяющего улучшать модели на децентрализованных данных без ущерба для приватности пользователей [20]. Архитектура будущих систем будет стремиться к большей унификации обработки разных типов контента и реализации принципов «модерации как кода», где политики безопасности могут гибко настраиваться и быстро развёртываться.

Несмотря на прогресс, полностью автономная модерация недостижима в обозримом будущем. Оптимальной представляется гибридная модель «человек в контуре» (human-in-the-loop), где ИИ выступает мощным ситом и ассистентом, а человек принимает окончательные решения в сложных этических и контекстных ситуациях. Это обеспечивает баланс между эффективностью, масштабируемостью и социальной ответственностью платформы.

Успешная система интеллектуальной модерации должна быть не просто технологическим решением, но и социально-технической системой, учитывающей культурный контекст, правовые нормы и этические принципы, оставаясь при этом технически эффективной и экономически целесообразной.

Литература:

1. Рассел С., Норвиг П. Искусственный интеллект: современный подход. 4-е изд. М.: Вильямс, 2021. 1408 с.

2. Гудфеллоу И., Бенджио Й., Курвилль А. Глубокое обучение. 2-е изд. М.: ДМК Пресс, 2018. 652 с.
3. Jurafsky D., Martin J. Speech and Language Processing. 3rd ed. Stanford: Stanford University Press, 2023. 624 p.
4. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998-6008.
5. Клеппман М. Проектирование интенсивно используемых приложений. СПб.: Питер, 2021. 560 с.
6. Ньюман С. Создание микросервисов. 2-е изд. СПб.: Питер, 2022. 512 с.
7. Huyen C. Designing Machine Learning Systems. Sebastopol: O'Reilly Media, 2022. 386 p.
8. Aggarwal C. Neural Networks and Deep Learning. Cham: Springer, 2018. 497 p.
9. Devlin J., Chang M.W., Lee K. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. 2019. P. 4171-4186.
10. Szeliski R. Computer Vision: Algorithms and Applications. 2nd ed. Cham: Springer, 2022. 947 p.
11. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. P. 770-778.
12. Tran D., Bourdev L., Fergus R. et al. Learning Spatiotemporal Features with 3D Convolutional Networks // Proceedings of the IEEE International Conference on Computer Vision. 2015. P. 4489-4497.
13. Rossler A., Cozzolino D., Verdoliva L. et al. FaceForensics++: Learning to Detect Manipulated Facial Images // Proceedings of the IEEE International Conference on Computer Vision. 2019. P. 1-11.
14. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 4th ed. Cambridge: Morgan Kaufmann, 2022. 752 p.

15. Kipf T.N., Welling M. Semi-Supervised Classification with Graph Convolutional Networks // International Conference on Learning Representations. 2017. P. 1-14.
16. Provost F., Fawcett T. Data Science for Business. Sebastopol: O'Reilly Media, 2013. 414 p.
17. Bishop C. Pattern Recognition and Machine Learning. Cham: Springer, 2006. 738 p.
18. Mittelstadt B., Russell C., Wachter S. Explaining Explanations in AI // Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019. P. 279-288.
19. Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier // Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135-1144.
20. McMahan B., Moore E., Ramage D. et al. Communication-Efficient Learning of Deep Networks from Decentralized Data // Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. 2017. P. 1273-1282.