DOI 10.37882/2223-2966.2025.09.30

МЕТОД ДВУХЭТАПНОЙ КЛАССИФИКАЦИИ

TWO-STAGE CLASSIFICATION METHOD

N. Teterin
T. Smolentseva

Summary. The article discusses a two-level classification method aimed at improving the quality of reliability of detecting destructive user behavior in social networks. The relevance of the study is due to the need to automate content moderation processes on the Internet, which will allow us to quickly respond to potentially destructive behavior. The purpose of this work is to develop a two-level classification method designed to detect destructive user behavior in social networks. The proposed approach includes two stages: at the first level, binary classification is performed to filter data, and at the second level, multi-class classification is used to accurately determine categories. The key quality criteria for the two-level classification method are: adaptability to new forms of threats, interpretability of decisions made, and compliance with expert assessment.

Keywords: two-stage classification, detection, destructive behavior, social networks, heterogeneous data.

Тетерин Николай Николаевич
Ассистент, РТУ МИРЭА (Москва)
teterin@mirea.ru
Смоленцева Татьяна Евгеньевна
Д.т.н., доцент, РТУ МИРЭА (Москва)
smoltan@bk.ru

Аннотация. В статье рассматривается метод двухэтапной классификации, направленный на повышение качества достоверности выявления деструктивного поведения пользователей в социальных сетях. Актуальность исследования обусловлена необходимостью автоматизации процессов модерации контента в сети интернет, что позволит оперативно реагировать на потенциально деструктивное поведение. Цель данной работы — разработка метода двухэтапной классификации, предназначенный для детекции деструктивного поведения пользователей в социальных сетях. Предложенный подход включает два этапа: на первом выполняется бинарная классификация для фильтрации данных, а на втором используется многоклассовая классификация для точного определения категорий. Ключевыми критериями качества метода двухуровневой классификации являются: адаптивность к новым формам угроз, интерпретируемость принимаемых решений и соответствие экспертной оценке.

Ключевые слова: двухэтапная классификация, детекция, деструктивное поведение, социальные сети, разнородные данные.

Современные социальные сети представляют собой сложную коммуникативную среду, где ежедневно появляются огромные объемы контента, значительную долю которого занимают материалы, содержащие деструктивное поведение [1]. Под сложной коммуникативной средой авторы подразумевают динамичное цифровое пространство с непрерывным потоком разноформатного контента и изменяющимися формами взаимодействия между пользователями [2]. Модерация такого контента становится все менее эффективной из-за масштабов и скорости распространения информации. В связи с этим возникает необходимость в разработке автоматизированных методов, способных не только обнаруживать деструктивный контент, но и классифицировать его по типам для последующего принятия управленческих решений [3].

Цель данной работы — разработка метода двухэтапной классификации, предназначенный для детекции деструктивного поведения пользователей в социальных сетях.

Анализ существующих исследований двухэтапных классификационных решений выявил существенные различия в технической реализации. В работе Лебедева И.С. предложен двухуровневый механизм обработки данных, где нижний уровень выполняет первичную классификацию, а верхний — динамически перераспределяет данные между моделями на основе изменений отношения между входными и выходными данными модели [4]. Масакулова Ж.А. в своем исследовании продемонстрировала альтернативный подход к технической реализации двухуровневой классификации, применив каскад из нейронных сетей для обработки многомерных временных рядов [5]. Однако существующие решения обладают рядом ограничений, в том числе они не ориентированы для работы с мультимодальными данными.

Метод двухэтапной классификации начинает свое применение после анализа мультимодальных данных, где обработанные данные структурированы и объединены [6]. Далее включается двухэтапный механизм, где первый уровень — это бинарная классификация, а второй уровень — это мультиклассовая классификация.

На этапе бинарной классификации система определяет содержит ли контент признаки деструктивного поведения, для этого используются алгоритмы машинного обучения, обученные на размеченных данных [7].

На этапе мультиклассовой классификации, если контент классифицирован как деструктивный, то система переходит к определению его категории. На данном эта-

пе используются более сложные модели искусственного интеллекта. Категории деструктивности могут включать: буллинг, экстремизм, троллинг и другие формы агрессии. Важной особенностью данного уровня является динамическое обновление категорий, что позволяет системе адаптироваться к новым видам угроз без необходимости полного переобучения [8].

Описание этапов метода двухэтапной классификации отображено на Рисунке 1.

После классификации система формирует решения, например отправка контента на дополнительную проверку модераторам. Для визуализации результатов используются отчеты и рекомендации, которые помогают модераторам принимать обоснованные решения [9].

На Рисунке 2 отражен процесс запуска классификации деструктивного контента. На данном этапе происходит загрузка модели Qwen2.5-7B, где обрабатывается 100 комментариев со скоростью 1.17 итераций в секунду.

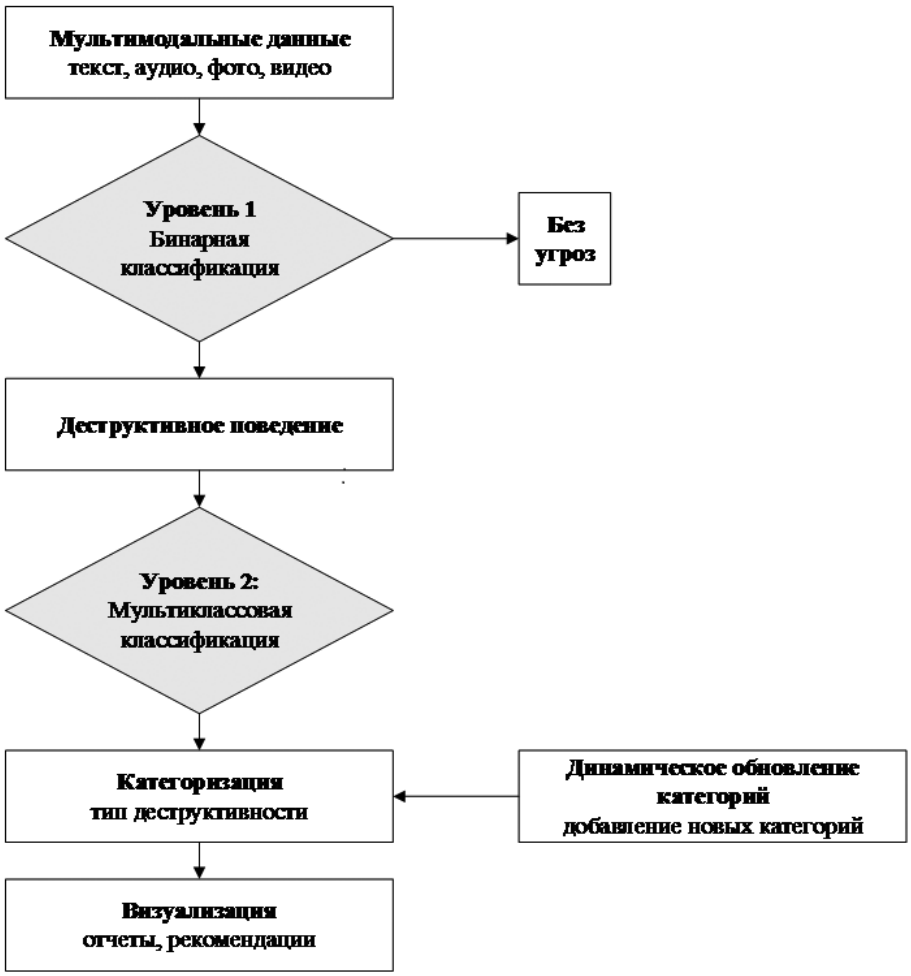

Рис. 1. Описание этапов метода двухэтапной классификации

```
Загрузка модели: Qwen/Qwen2.5-7B-Instruct (с 4-битной квантизацией)
Используемое устройство: cuda
Sliding Window Attention is enabled but not implemented for `sdpa`; unexpected results may be encountered.
Loading checkpoint shards: 100% ████████████████████ 4/4 [00:16<00:00,  4.15s/it]
4-битная модель успешно загружена на GPU.
Модель и токенизатор готовы.
Загрузка данных из файла: data/vk_comments_podslushka_mirea1.xlsx
Найдено комментариев для анализа: 1119
Сколько комментариев вы хотите обработать? (Введите число от 1 до 1119 или 'все'):  100
Будет обработано комментариев: 100
Начало классификации комментариев...
Классификация:   8% ██                  8/100 [00:07<01:18,  1.17it/s]
```

Рис. 2. Процесс запуска классификации

На Рисунке 3 интерфейс системы предлагает три варианта анализа: только текст, только медиа или комплексный, а именно медиа и текст.

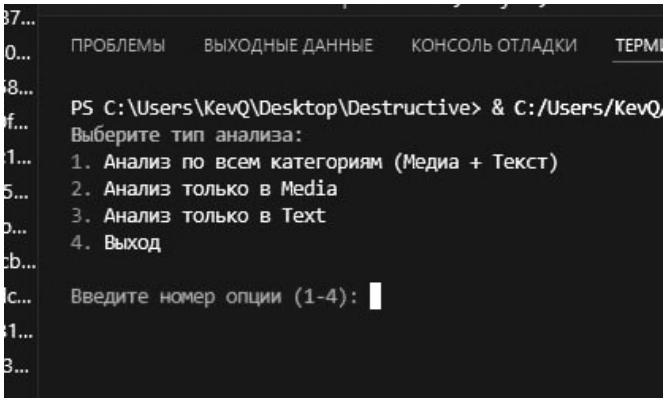

Рис. 3. Анализ контента по типу медиа и текст

На Рисунке 4 отображен перечень категорий деструктивного контента, который определяет тип деструктивности на этапе мультиклассовой классификации.

Важно отметить, что данная классификация не является статичной и включает возможность динамического обновления категорий деструктивного поведения в процессе работы.

Описанный выше метод двухэтапной классификации опирается на три ключевых критерия качества достоверности: адаптивность, интерпретируемость и соответствие экспертной оценке. Данные критерии обеспечивают не только высокую точность работы системы, но и ее практическую применимость в условиях динамично изменяющегося контента социальных сетей [10].

На Рисунке 5 представлены ключевые критерии качества по показателю достоверность.

Адаптивность — способность системы динамически обновлять категории классификации при появлении новых форм деструктивного контента без полного переобучения. Измеряется по метрикам: F1-score и FPR.

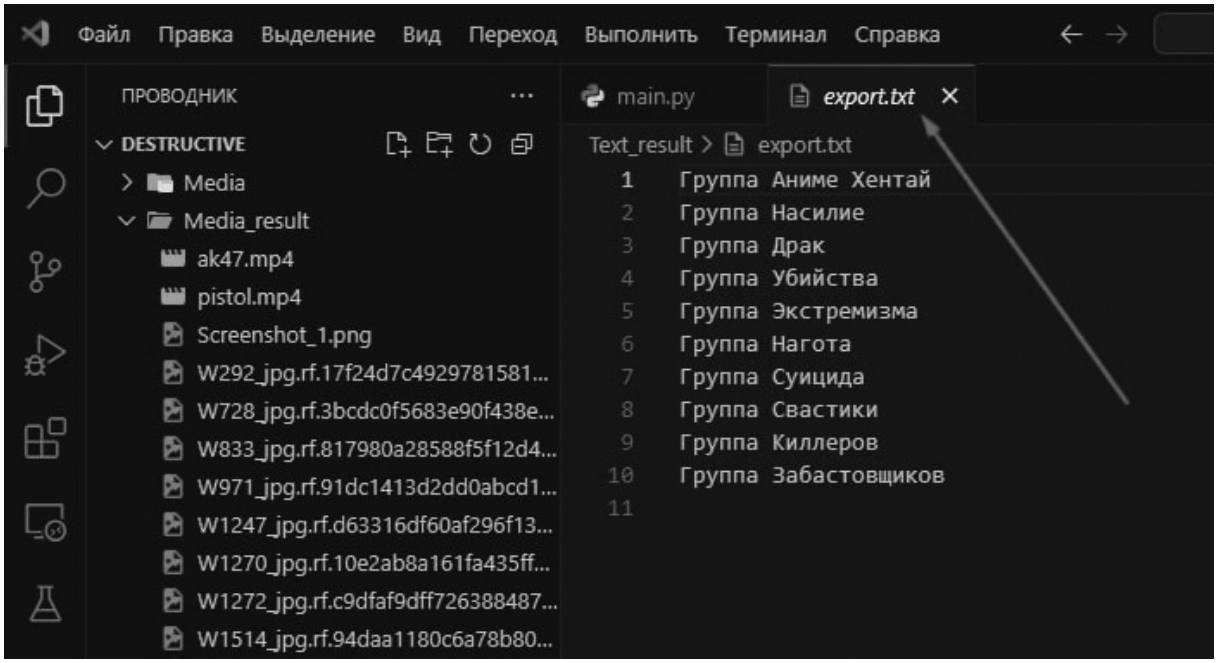

Рис. 4. Категории деструктивного поведения

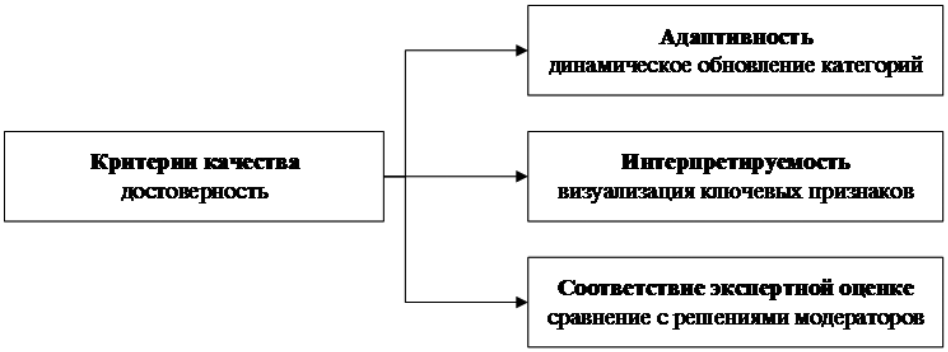

Рис. 5. Критерии качества по показателю достоверность

F1-score — метрика, которая показывает оценку эффективности бинарной классификации от 0 до 1.

FPR (false positive rate) — метрика, которая показывает как часто классификатор определяет объекты в неверный класс.

Интерпретируемость — возможность объяснить, на основе каких признаков система приняла решение. Измеряется по использованию XAI методов: LIME и SHAP.

XAI — это методы и инструменты, которые делают искусственный интеллект более понятным для людей.

LIME — метод создания более простой модели для локальной аппроксимации классификатора или регрессора.

SHAP — метод присвоения баллов важности каждому признаку для конкретного случая, объясняя конкретный прогноз.

Соответствие экспертной оценке — степень согласованности решений между системой и модератором. Измеряется по метрике: Cohen's Kappa.

Cohen's Kappa — статистическая метрика, которая измеряет уровень согласия между двумя экспертами.

В таблице 1 показаны ключевые критерии достоверности и их метрики.

Таблица 1.

Критерии достоверности и их метрики измерения

Критерий	Метрики измерения
Адаптивность	F1-score, FPR
Интерпретируемость	% объяснимых решений (SHAP/LIME)
Соответствие экспертной оценке	Cohen's Kappa

Вышеописанные метрики позволяют объективно оценить качество метода двухуровневой классификации и его применимость в реальных условиях модерации.

Таким образом, вышеописанный метод двухуровневой классификации демонстрирует высокую необходимость применения в задачах детекции деструктивного поведения пользователей в социальных сетях.

Использование приведенных в работе метрик обеспечивает достоверную оценку качества метода двухуровневой классификации. Адаптивность гарантирует устойчивость к новым угрозам, интерпретируемость — прозрачность, а соответствие экспертной оценке — практическую применимость.

Перспективы дальнейших исследований связаны с совершенствованием алгоритмов классификации и разработкой механизмов для более точного определения новых форм деструктивности.

ЛИТЕРАТУРА

1. Гусев, М.М. Исследование зависимости влияния соционического типа пользователя социальной сети на его поведение в социальной сети / М.М. Гусев, А.Н. Гусева, Т.В. Кораблина // Вестник Сибирского государственного индустриального университета. — 2020. — № 2(32). — С. 71–73. — EDN SQXBTW.
2. Тетерин, Н.Н. Общие вопросы анализа деструктивного поведения пользователей в социальных сетях / Н.Н. Тетерин // Актуальные проблемы деятельности подразделений уголовно-исполнительной системы: Сборник материалов Всероссийской научно-практической конференции. В 3-х томах, Воронеж, 24 октября 2024 года. — Воронеж: ООО Издательско-полиграфический центр «Научная книга», 2024. — С. 96–99. — EDN OWDJUT.
3. Тетерин, Н.Н. Концептуальный подход классификации деструктивного поведения с применением технологий искусственного интеллекта / Н.Н. Тетерин, В.В. Смоленцева // Актуальные проблемы прикладной математики, информатики и механики: Сборник трудов Международной научной конференции, Воронеж, 02–04 декабря 2024 года. — Воронеж: Научно-исследовательские публикации, 2025. — С. 326–329. — EDN DZEHAU.
4. Тетерин, Н.Н. К вопросу формализации задачи выявления деструктивного поведения с применением технологий искусственного интеллекта / Н.Н. Тетерин, В.В. Смоленцева // Тенденции развития науки и образования. — 2024. — № 114-10. — С. 81–83. — DOI 10.18411/trnio-10-2024-440. — EDN EEBVGI.
5. Лебедев, И.С. Применение многоуровневых моделей в задачах классификации и регрессионного анализа / И.С. Лебедев // Информатика и автоматизация. — 2023. — Т. 22, № 3. — С. 487–510. — DOI 10.15622/ia.22.3.1. — EDN TUABWM.
6. Мусакулова, Ж.А. Применение двухуровневой нейронной сети Кохонена в медицинской задаче классификации данных / Ж.А. Мусакулова // Евразийское Научное Объединение. — 2021. — № 2-2(72). — С. 98–102. — DOI 10.5281/zenodo.4599668. — EDN PXTUHG.
7. Павлюченко, М.В. Анализ зависимости точности бинарной классификации текстов от применения мета-функций для различных алгоритмов классификации / М.В. Павлюченко, Т.В. Кабанова // Математическое и программное обеспечение информационных, технических и экономических систем : Материалы VIII Международной молодежной научной конференции, Томск, 26–30 мая 2021 года / Под общей редакцией И.С. Шмырина. Том 306. — Томск: Национальный исследовательский Томский государственный университет, 2021. — С. 42–47. — DOI 10.17223/978-5-907442-42-9-2021-8. — EDN EIEQNU.
8. Стужук, И.Г. Анализ степени деструктивности аккаунтов в социальных сетях / И.Г. Стужук // Право и общество. — 2022. — № 4(9). — С. 67–70. — EDN JTFEIX.
9. Эргешова, А.Ж. Роль искусственного интеллекта в профилактике деструктивного поведения среди молодежи / А.Ж. Эргешова // Актуальные тенденции социальных коммуникаций: история и современность: Сборник научных статей. — Ижевск: Издательский дом «Удмуртский университет», 2024. — С. 439–442. — EDN IWVBYH.
10. Шевченко, Д.А. Актуальные аспекты воздействия различных информационных источников на деструктивное поведение молодежи / Д.А. Шевченко // Актуальные вопросы юридической науки глазами молодых исследователей: Сборник статей по итогам Четвертой Всероссийской научной конференции курсантов, студентов, адъюнктов, аспирантов и соискателей, Рязань, 02 февраля 2024 года. — Москва — Нижний Новгород: Постер-М, Российская академия народного хозяйства и государственной службы при Президенте РФ, 2024. — С. 155–158. — EDN KXQXNK.