Классификация токсичности текстов

Русанов Дмитрий

ВШЭ СПб 2025

4 декабря 2025 г.

Цель работы

Постановка задачи

Задача бинарной классификации: определение токсичности русскоязычных текстов (комментариев) с получением калиброванных вероятностных оценок.

Задачи, что хотим решить:

- **Классификация:** автоматически определять, является ли комментарий токсичным
- **Калибровка:** получать надёжные вероятности (хотим, чтобы если модель даёт X%, то в X% случаев это действительно токсичный текст)

Практическая применимость:

- Автоматическая модерация контента
- Фильтрация токсичных комментариев в реальном времени
- Помощь модераторам в приоритизации проверки сообщений

Исходные данные

Источник: публичные датасеты с HuggingFace, объединённый корпус для обучения и анализа OOF-предсказаний.

Объём и структура:

- **N = 251,317** примеров в итоговом объединённом корпусе (full combined)
- Каждый пример: `text`, `label` (0/1), `soft_label` (OOF-вероятность класса 1)
- OOF-предсказания получены через StratifiedKFold (5 фолдов) и используются для оценки калибровки

Ключевые характеристики

Разнообразие источников:

- AlexSham — общие русскоязычные комментарии
- toxic_dvach — форумные обсуждения

Особенности датасета:

- **Дисбаланс классов:** токсичных комментариев обычно меньше (∼20-40%)
- **Вариативность длины:** от коротких фраз (5-10 слов) до длинных текстов (100+ слов)
- **Контекстная зависимость:** некоторые слова токсичны только в определённом контексте

Предобработка:

- Дедупликация
- Нормализация
- Генерация OOF soft labels через StratifiedKFold для анализа калибровки

Используемые алгоритмы

Выбранный подход: TF-IDF Vectorization + Logistic Regression

1. Векторизация (TF-IDF):

- **Что делаем:** преобразует текст в числовой вектор
- max_features = 10 000
- ngram_range = (1, 2)
- Учитывает важность слов и биграмм

2. Классификатор (LogReg):

- **Что делаем:** линейная классификация
- Параметр регуляризации C = 1.0
- Solver: 'lbfgs' (эффективен для средних данных)
- Выдаёт калиброванные вероятности

3. Калибровка (CalibratedClassifierCV):

- **Метод:** Platt scaling (method='sigmoid')
- **Валидация:** StratifiedKFold (cv=5)
- **Цель:** улучшить соответствие предсказанных вероятностей реальным частотам

Обоснование выбора

TF-IDF + Logistic Regression

- **Интерпретируемость:** веса признаков дают прозрачную локальную трактовку решений.
- **Быстрота и экономичность:** обучение и предсказание выполняются быстро на CPU, подходит для экспериментов и демонстраций.
- **Стабильность:** линейная модель менее склонна к переобучению на шумных метках при базовой регуляризации.
- **Простая калибровка:** комбинируется с Platt scaling (CalibratedClassifierCV) для получения надёжных вероятностей.
- **Удобный baseline:** лёгкая репликация результатов и быстрый отклик при изменении предобработки/фичей.

Как измерялось качество моделей

Метод валидации: StratifiedKFold (5 фолдов)

- Данные разбиваются на 5 частей с сохранением распределения классов
- Модель обучается на 4 частях, тестируется на 1
- Процесс повторяется 5 раз (каждая часть один раз в тестовой выборке)
- Результат: OOF (Out-Of-Fold) предсказания для всего датасета

Используем стандартные метрики классификации:

- **Accuracy** — доля правильных ответов
- **Precision** — из предсказанных токсичных, сколько действительно токсичны
- **Recall (Sensitivity)** — из всех токсичных, сколько модель нашла
- **F1-Score** — гармоническое среднее Precision и Recall
- **ROC AUC** — площадь под ROC-кривой

Метрики калибровки вероятностей

Метрики для оценки калибровки:

- **Brier Score:** $BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2$
  - Среднеквадратичная ошибка между предсказанными вероятностями и истинными метками
  - Диапазон: [0, 1], чем ниже — тем лучше
- **Reliability Diagram (Calibration Plot):**
  - Визуализация: предсказанные вероятности vs. реальные частоты
  - Разбиваем вероятности на бины
  - Для каждого бина считаем среднюю предсказанную вероятность и реальную долю положительных примеров

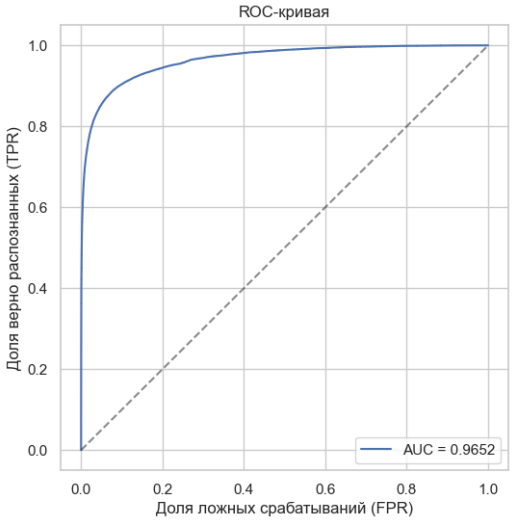
Итоговые результаты

Метрики на OOF-данных (5-fold cross-validation):

Метрика	Значение
Accuracy	0.9409
Precision	0.9022
Recall	0.7710
F1-Score	0.8315
ROC AUC	0.9652
Brier Score	0.0455

**Метрики рассчитаны на OOF-предсказаниях (5 фолдов) combined_oof_full.csv* (полный combined). N = 251,317 примеров.

ROC-кривая


ROC-кривая
Доля верно распознанных (TPR)
Доля ложных срабатываний (FPR)
AUC = 0.9652
0.0
0.2
0.4
0.6
0.8
1.0

Precision–Recall кривая

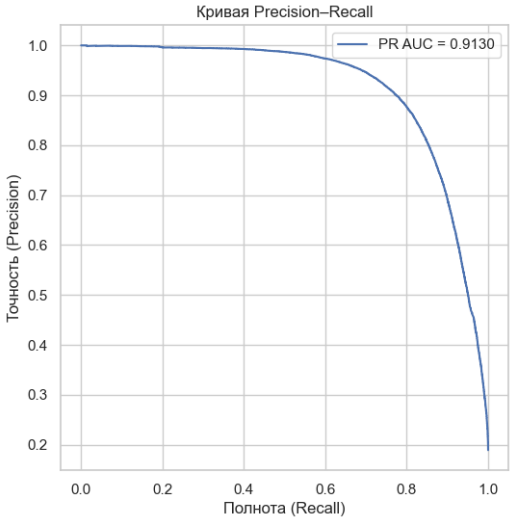
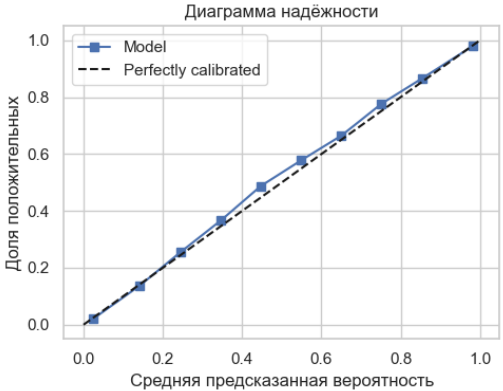
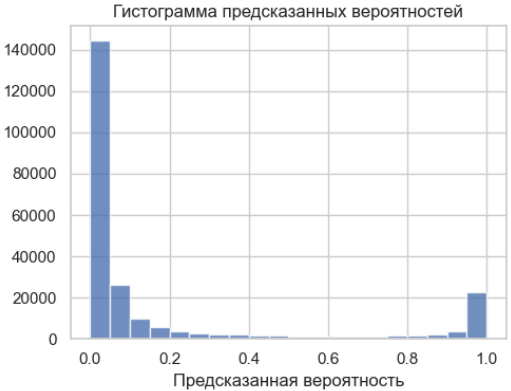


Диаграмма калибровки и гистограмма предсказанных вероятностей





Матрица ошибок