Классификация комментариев на ревью кода

Результаты

Модель	Accuracy	Precision	Recall	F1-Score
Logistic Regression + CountVectorizer	0.8931	0.7111	0.7612	0.7350
Logistic Regression + TF-IDF	0.8811	0.6605	0.8012	0.7239
Random Forest + CountVectorizer	0.8625	0.6295	0.7162	0.6699
Random Forest + TF-IDF	0.8528	0.6031	0.7171	0.6550
RoBERTa	0.9198	0.8157	0.7624	0.7881
CodeBERT	**0.9229**	**0.8228**	**0.7723**	**0.7967**

Проблемы и их решения

Дисбаланс классов

Проблема: Датасет несбалансирован — токсичных комментариев всего 20.3%.

Решение:

- Использована стратифицированная выборка при разбиении train/test
- Рассмотрена возможность применения class_weight='balanced' в Logistic Regression
- Для трансформеров: мониторинг метрик на каждом классе отдельно

Вычислительные ресурсы

Проблема: Обучение трансформеров требует GPU.

Решение:

- Использовался Google Colab с бесплатным GPU (Tesla T4)
- Включена mixed precision training (fp16=True) для ускорения
- Уменьшен max_length до 128 токенов (вместо 512)
- Batch size ограничен 16 для экономии памяти

Долгое обучение трансформеров

Облачные ресурсы в google colab все равно имеют малую производительность и лимиты по использованию
Решение:

- Использовал early stopping (patience=2)

Очистка датасета связанная с чувствительной лексикой

Словарь который был представлен в репо ToxicCR имел синтаксические ошибки, в строках
лежали просто объект str, когда на самом деле это было регулярное выражение (r-строка в python)

Решение:

- Отформатировать исходную структуру данных, было принято решение этого не делать.
- Потенциально это может повысить точность моделей.

Итог

Дальнейшие улучшения:

1. Собрать больше данных, особенно токсичных комментариев
2. Использовать аугментацию данных (back translation, paraphrasing)
3. Ensemble методы: комбинировать предсказания нескольких моделей
4. Fine-tuning трансформеров дольше (5-10 эпох) с early stopping
5. Гиперпараметрическая оптимизация (Grid Search, Bayesian Optimization)

Все графики располагаются в директориях `4_results`, `3_transformers`, `2_classical_models`.
