1. Введение

Проект "Тон - Корректор" направлен на создание инструмента для автоматического выявления токсичных высказываний в тексте с последующим их визуальным выделением. Актуальность задачи обусловлена растущей потребностью в фильтрации агрессивного, оскорбительного и деструктивного контента в социальных сетях, мессенджерах и других цифровых платформах.

Основная цель — разработка модели классификации текста по тональности, способной эффективно определять токсичность с высокой точностью. Для этого были протестированы три модели машинного обучения: **FastText**, **ru-BERT** и **FRIDA**, каждая из которых обладает уникальными особенностями и подходами к обработке естественного языка.

2. Датасеты и предобработка данных

Для обучения и оценки моделей использовались следующие источники данных:

- **Основной датасет**: https://www.kaggle.com/datasets/blackmoon/russian-language-toxic-comments.
- **Дополнительные данные** — 20 000 текстов (10 000 с меткой "токсично" и 10 000 с "нетоксично"), добавлены ввиду малости первого датасета. Взяты отсюда: https://huggingface.co/datasets/AlexSham/Toxic_Russian_Comments

Предобработка включала:

- Очистку текстов от спецсимволов, HTML-тегов смайлов и дубликатов.
- Лемматизацию и приведение слов к нижнему регистру (для FastText и ru-BERT).
- Разделение данных на обучающую, валидационную и тестовую выборки (стандартное соотношение 80/10/10).

3. Описание моделей и их настройка

3.1. FastText

FastText — это библиотека для эффективного обучения классификации текстов и создания векторных представлений слов, разработанная исследователями из Facebook AI Research (FAIR).

Преимущества:

- Эффективность при работе с редкими словами и опечатками благодаря анализу подстрок.
- Высокая скорость обучения и предсказания по сравнению с нейросетевыми подходами.
- Поддержка многоклассовой классификации.

Настройка:

- Использовались предобученные векторы для русского языка.
- Оптимизация гиперпараметров (размер n-грамм, скорость обучения, размер эмбеддингов).

3.2. ru-BERT

ru-BERT — адаптированная версия BERT для русского языка (blanchefort/rubert-base-cased-sentiment), предобученная на 300 000+ текстах с метками тональности (POSITIVE, NEGATIVE, NEUTRAL).

Fine-tuning:

- Заморозка части слоев для сохранения языковых знаний.
- Добавление классификационной головы под задачу бинарной классификации (токсично/нетоксично).

3.3. FRIDA

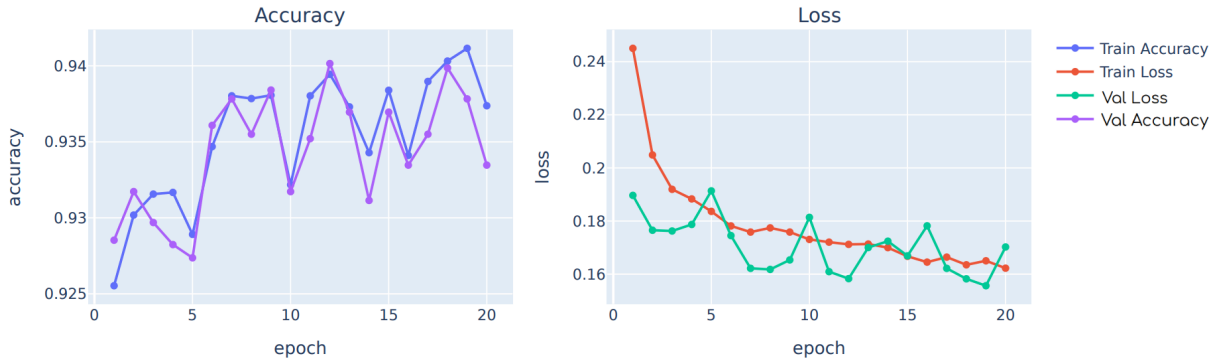
FRIDA — T5-based трансформер, лидирующий в рейтинге русскоязычной версии бенчмарка MTEB.

Особенности:

- Поддержка множества NLP-задач, включая классификацию токсичности (префикс "categorize_sentiment: ").
- Высокое качество эмбеддингов, что особенно полезно для задач семантического анализа.

Fine-tuning:

- Добавление линейного классификатора поверх замороженного эмбеддера.
- Обучение только головы модели для ускорения процесса.
- Графики обучения показали рост accuracy и снижение loss (можно посмотреть ниже)



4. Результаты и сравнение моделей

	accuracy	precision	recall
FastText	0.8314	0.8694	0.7173
ru-Bert	0.7108	0.6267	0.8116
FRIDA	0.9340	0.9182	0.9300

Анализ:

- **FRIDA** показала наилучшие результаты по всем метрикам, демонстрируя сбалансированность между точностью и полнотой.
- **FastText** уступил в recall, но достиг высокой precision, что полезно для минимизации ложных срабатываний.
- **ru-BERT** показал низкую точность, вероятно, из-за неполного соответствия предобученных меток (нейтральный/негативный тон ≠ токсичность).

5. Выводы и рекомендации

1. **FRIDA** — оптимальный выбор для внедрения в приложение благодаря высокой точности и адаптивности к разным типам текстов.
2. **FastText** может быть полезен в сценариях, где важна скорость обработки (например, модерация в реальном времени).
3. **ru-BERT** требует дополнительной доработки:
    - Расширение датасета с акцентом на токсичные примеры.
    - Эксперименты с разными архитектурами классификационной головы.

Дальнейшие шаги:

- Разработка API для интеграции модели в веб- и мобильные приложения.
- Добавление контекстного анализа (учет сарказма, иронии).
- Создание пользовательского интерфейса с подсветкой токсичных фрагментов.