Детоксификация русскоязычных текстов

Ольга Пичужкина, Татьяна Гнедина

Данные

https://www.kaggle.com/blackmoon/russian-language-toxic-comments
https://www.kaggle.com/alexandersemiletov/toxic-russian-comments
Всего: 262702 комментария (источники 2ch.hk, pikabu.ru, ok.ru), из них 213271 не-токсичных и 49431 токсичных.

Обучающая выборка: 213271 не-токсичных комментария, 37073 токсичных;
Обучающая выборка: параллельный корпус из 500 комментариев;
Тестовая выборка: 12358 токсичных комментария.

Baselines

- **Duplicate**: “наивный бейзлайн” — просто продублировать предложение без изменений;
- **Delete**: удалить из предложения все слова, входящие в словарь токсичных слов, а также все слова, леммы которых совпадают с леммами слов из словаря.

Метрики оценки

- **style transfer accuracy (STA)** — бинарная метрика стиля, рассчитываемая с помощью предобученного классификатора токсичности на основе BERT;
- **cosine similarity (CS)** — метрика косинусной близости, рассчитываемая с помощью эмбеддингов FastText;
- **Fluency score (FL)** — метрика естественности, рассчитываемая с помощью предобученного классификатора искажённости текста на основе BERT

- **Joint score (JS) — STA * CS * FL**

Бейзлайны: метрики оценки

Method	STA↑	CS↑	FL↑	JS↑
Baselines				
Duplicate	0.07	1.00	1.00	**0.06**
Delete	0.25	0.96	0.89	**0.23**

Модель

1. **Toxic word detection** — we train a binary classifier (LogReg) to detect toxic words;
2. **Toxic word replacement** — to replace words classified as toxic, we use one of pre-trained NLP models for Russian language (either **ruBERT-large** or **ruRoBERTa-large**). From the top-10 of model predictions we select one that is 1) non-toxic 2) closest to the original word (word embeddings are generated with the FastText model);
3. **Toxic word deletion** — if a non-toxic replacement wasn't found in the top-10 of model predictions, we delete the word.