Детоксификация русскоязычных текстов

Ольга Пичужкина, Татьяна Гнедина

Задача

Написать модель для задачи автоматической детоксификации русскоязычных текстов.

Актуальность/мотивация: автоматическая детоксификация текстов может быть использована в социальных сетях.

Команда и роли

Ольга Пичужкина

- Сбор и препроцессинг данных
- Обучение моделей на не-параллельном корпусе комментариев
- Оценка качества моделей

Татьяна Гнедина

- Сбор и препроцессинг данных (параллельный корпус комментариев)
- Обучение моделей на параллельном корпусе комментариев

Данные

https://www.kaggle.com/blackmoon/russian-language-toxic-comments

https://www.kaggle.com/alexandersemiletov/toxic-russian-comments

Всего: 262702 комментария, из них 213271 не-токсичных и 49431 токсичных.

Также для обучения моделей, требующих параллельного корпуса данных, мы создадим вручную параллельный корпус из 500 комментариев (комментарий + его детоксифицированная версия).

Baselines

- **Duplicate**: “наивный бейзлайн” — просто продублировать предложение без изменений;
- **Delete**: removal of rude and toxic from pre-defined vocab;
- **Retrieve**: retrieval based on cosine similarity between word embeddings from non-toxic part of RuToxic dataset;

Метрика оценки качества

Геометрическое среднее трёх метрик:

- **style transfer accuracy (STA)** — берём предобученный классификатор токсичных/нетоксичных комментариев, и им проверяем, сколько из прошедших через модель комментариев получились не-токсичными (считаем accuracy)
- **cosine similarity (CS)** — косинусная близость между векторными эмбеддингами оригинального комментария и детоксифицированного (насколько новый комментарий похож по смыслу на оригинальный)
- **1 / perplexity (PPL)** — насколько хорошо предобученная языковая модель (мы возьмём ruGPT2Large) может предсказать получившийся комментарий

План действий

- Собрать датасет комментариев на русском языке, размеченных как токсичные/нетоксичные
- Из полученного датасета выбрать 500 токсичных комментариев и вручную создать параллельный датасет комментариев
- Fine-tuning **ruGPT** на параллельном датасете
- Fine-tuning **conditional BERT** на не-параллельном датасете
- CNN на не-параллельном датасете
- Сравнить качество моделей

Список использованной литературы

- Jin et al. Deep Learning for Text Style Transfer: A Survey
- Dementieva et al. Methods for Detoxification of Texts for the Russian Language. Диалог-2021 (https://github.com/skoltech-nlp/rudetoxifier) — отсюда взяты бейзлайны и метрики оценки
- https://github.com/sberbank-ai/ru-gpts — отсюда взята модель для обучения на параллельных данных
- https://huggingface.co/DeepPavlov/rubert-base-cased-conversational и
- https://huggingface.co/Geotrend/bert-base-ru-cased — отсюда взяты модели для обучения на не-параллельных данных
- https://huggingface.co/sismetanin/rubert-toxic-pikabu-2ch — отсюда взят предобученный классификатор токсичных комментариев, который мы будем использовать для подсчёта style text transfer
- https://github.com/sberbank-ai/ru-gpts#Pretraining-ruGPT2Large — отсюда взята модель, которую мы будем использовать для подсчёта perplexity