

DIALOGUE 2022

RUSSIAN TEXT DETOXIFICATION BASED ON

PARALLEL CORPORA

РОМАН КАЗАКОВ, КСЕНИЯ ПЕТУХОВА, ВЕРОНИКА СМЛГА

БКЛ182

Дано: датасет токсичных комментариев

Задача: нужно привести комментарии к нейтральному стилю

АКТУАЛЬНОСТЬ

- ▶ Социальная значимость: с ростом активности в социальных сетях распространяются такие явления, как травля и буллинг; с помощью систем детоксификации можно попытаться обезопасить общение незнакомых людей в чатах и комментариях
- ▶ Научная значимость: перенос стиля текста – комплексная задача NLP, задействующая разные уровни языка; качественное решение такой задачи стало возможным лишь с появлением сложных нейросетевых архитектур (например, CNN)
- ▶ Бизнес значимость: решение этой задачи может упростить модерацию социальных сетей, а также использоваться для контроля ответов диалоговых систем (чат-ботов и голосовых ассистентов)
- ▶ *Творческая значимость: кажется, что путей решения этой задачи может быть очень много

ОБЗОР ЛИТЕРАТУРЫ

- ▶ [Dale et al., 2021] – в работе представлены две модели для решения задачи детоксификации: 1) ParaGeDi: T5 для парафразы + GPT, чтобы выбирать самое вероятное, но при этом не токсичное слово; 2) CondBERT: детекция токсичных слов и подстановка вместо него слова, близкого по семантике
- ▶ [Dementieva et al., 2021] – первая работа, посвященная автоматической детоксификации русских текстов, в ней также представлены два подхода: 1) supervised подход на основе ruGPT-3 (команда "Перепарафразируй _ >>>"); 2) CondBERT, как в [Dale et al., 2021]

ДАННЫЕ

- ▶ **Параллельный датасет:** токсичное предложение на русском языке и 1-3 его нетоксичных аналога
- ▶ **Сообщения из соцсетей:** Одноклассники, Пикабу и Твиттер

Data	Кол-во токсичных предложений
Train	3539
Development	800
Test	875

МЕТРИКИ ОЦЕНКИ

- ▶ **Style transfer accuracy (STA)**: бинарная метрика стиля, рассчитываемая с помощью классификатора токсичности на основе BERT, обученного на датасете русскоязычных токсичных комментариев (насколько удалось детоксифицировать)
- ▶ **Meaning preservation score (SIM)**: метрика косинусной близости, рассчитываемая с помощью эмбедингов предложений LaBSE (насколько порождённое предложение сохранило семантику)
- ▶ **Fluency score (FL)**: метрика естественности, рассчитываемая на основе классификатора BERT, обученного на русскоязычных комментариях из социальных сетей и их автоматически сгенерированных аналогах (насколько текст похож на порождённый носителем языка)
- ▶ **Joint score (J)**: $STA \cdot SIM \cdot FL$
- ▶ Для финальной оценки private теста будет использоваться ручной аналог каждой из этих метрик: бинарная STA, бинарная SIM и трехклассовая FL

BASELINE

- **Delete-base:** удаление токсичных слов из предложений (на основе словаря). Наиболее примитивный вариант, не требует наличия параллельных корпусов

STA	0.53
SIM	0.87
FL	0.82
J	0.36

BASELINE

- ▶ **T5-base**: на трейне обучается модель Сбера ruT5-base. Наиболее стандартное решение при наличии параллельных данных, использует современную контекстуальную модель

STA	0.74
SIM	0.76
FL	0.9
J	0.45

ЧТО СДЕЛАЛИ МЫ

1. ОПРЕДЕЛЕНИЕ ТОКСИЧНЫХ СЛОВ. БЫЛО

Бинарный классификатор: пока что логистическая регрессия

Фичи: fastText эмбединги (базовая модель для РЯ)

Данные: словарь, предоставленный организаторами (10854 уникальные леммы), и обучающий набор

1. ОПРЕДЕЛЕНИЕ ТОКСИЧНЫХ СЛОВ. СТАЛО

Бинарный классификатор (seq-to-seq): fine-tuned ruBERT-tiny

Данные: словарь, предоставленный организаторами (10854 уникальные леммы), и обучающий набор. Обучающий набор размечен на предмет токсичности каждого слова (0/1) в зависимости от наличия слова в словаре.

Сложности: ВРЕ-токенизация и добавление специальных токенов BERT-a приводит к мисмэтчу между получившимися токенами и оставшимися с предыдущего этапа лейблами; это решается с помощью кастомной функции и метода Datasets.map (HuggingFace).

2. ПОДБОР КАНДИДАТОВ ДЛЯ ЗАМЕНЫ

- ▶ Для токенов, которые наша модель классифицировала как токсичные, находили замену с помощью модели ruRoBERTa-large (Сбер)
- ▶ Чтобы выбрать лучшего кандидата:
 1. смотрели топ-30 (раньше топ-10) подстановок, предлагаемых моделью
 2. проверяли их наличие в словаре токсичной лексики
 3. загружали предложение с подстановкой в классификатор и смотрели на её лейбл
 4. если подстановка не токсичная, подставляли её, если нет, то шли дальше по списку с вероятностями подстановок

3. УДАЛЕНИЕ НЕИЗМЕНЕННЫХ ТОКСИЧНЫХ СЛОВ

- ▶ В случаях, когда классификатор маркировал токен как токсичный, но ruRoBERTa не находила нетоксичных подстановок, на которые этот токен можно заменить, мы удаляли токсичное слово

МЕТРИКИ НА TEST

Метрика	Было	Стало
STA	0.55	0.64
SIM	0.87	0.89
FL	0.89	0.92
J	0.4	0.51

ПРИМЕРЫ

- ▶ В такой ситуации виноваты **п*дарасы** из Ростелекома у которых даже кошка может купить фильм с пульта. **Дебилы бл*ть.** → В такой ситуации виноваты **люди** из Ростелекома у которых даже кошка может купить фильм с пульта. **Тьфу короче.**
- ▶ А **нахрена** тогда ты здесь это писал? → А **зачем** тогда ты здесь это писал ?
- ▶ а судью **п*дора** бог накажет. все возвращается бумерангом. → а судью пусть бог накажет. все возвращается бумерангом.
- ▶ и как земля таких **уродов** носит? → и как земля таких **людей** носит?
- ▶ где они **бл*ди** такие зарплаты нашли? → где они там такие зарплаты нашли?
- ▶ только тебя **д*лбо*ба** там нехватало. → только тебя мне там нехватало.

ЛИТЕРАТУРА

- ▶ Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova and Rada Mihalcea. "Deep Learning for Text Style Transfer: A Survey." ArXiv abs/2011.00416 (2020)
- ▶ Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. "Methods for Detoxification of Texts for the Russian Language" Multimodal Technologies and Interaction 5 (2021): no. 9: 54. <https://doi.org/10.3390/mti5090054>
- ▶ David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov and Alexander Panchenko. "Text Detoxification using Large Pre-trained Neural Models." EMNLP (2021)

МЫ

- ▶ Рома: идентификация токсичных слов
- ▶ Ксюша: замена слов с помощью ruRoBERTa
- ▶ Ника: удаление незамененных токенов

Спасибо за внимание!