# DIALOGUE 2022
# RUSSIAN TEXT DETOXIFICATION BASED ON PARALLEL CORPORA

РОМАН КАЗАКОВ, КСЕНИЯ ПЕТУХОВА, ВЕРОНИКА СМИЛГА

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 el al., 2021] – первая работа, посвященная автоматической детоксификации русских текстов, в ней также представлены два подхода: 1) supervised подход на основе ruGPT-3 (команда "Перефразируй _ >>>"); 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 | 1474 |

# 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 эмбеддингов 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

Бинарный классификатор: *пока что* логистическая регрессия

Фичи: fastText эмбеддинги (базовая модель для РЯ)

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

Метрики: *precision* = 0,88, *recall* = 0,88, *F1-score* = 0,88

# 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, мы смотрели на топ-10 вариантов замены, предлагаемых моделью, и оценивали косинусную близость между заменяемым словом и кандидатом с помощью векторов fastText

# 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 DEV

| | |
|---|---|
| STA | 0.55 |
| SIM | 0.87 |
| FL | 0.89 |
| J | 0.4 |

# ПРИМЕРЫ

- ***пизд*ж**! температуры горения хватит чтобы её расплавить к **х*рам**.. -> Только температуры горения хватит чтобы её расплавить к **нулю***
- ***пизд*бол** ---- а сам где кормишься ? - > а сам где кормишься*
- *этому сайту я давно не доверяю, пишут разную **х**... -> этому сайту я давно не доверяю, пишут разную **чушь***
- *это что за **высер придурка обиженного**🤪 -> это что за **позиция***
- *такой **п*зды** ни одна басня ещё не удостоилась. -> такой **участи** ни одна басня ещё не удостоилась.*

# ПЛАНЫ

- Заменить бинарный классификатор (логистическую регрессию) на несколько CNN-слоев. Кажется, они неплохо работают на таких задачах
- Fine-tuning ruRoBERTa: *здесь хотелось бы помощи*
- Улучшить процесс выбора операции (замена / удаление), *может быть тоже можно как-то попробовать с помощью нейросетей?*
- *(Возможно)* Имплементировать синтаксические ϕ

# 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!