детоксификация текстов

В. Ганеева,
Я. Лабенская

Детоксификация текстов

- Люди в интернете матерятся и оскорбляют других людей. По разным соображениям это может быть неприемлимым для части пользователей интернета (например, для детей). И эта часть достаточно большая. Запрет и фильтры, которые удаляют, например, комментарии по ключевым словам – это не очень хорошо, да и с задачей они справляются не полностью, так как оскорбления – это не только закрытый список слов
- Что такое детоксификация? Перефразирование оскорбляющих комментариев и/или удаление из них оскорблений
- Задача – создать или дообучить модель, которая была бы не слишком большой и справлялась с заданием лучше, чем удаление оскорблений по словарю

Актуальность

Практические применения:

- Вам не хочется видеть мат и оскорбления? Прекрасно, вот расширение для браузера – и вы их не увидите
- Беспокоитесь за словарный запас ребёнка, который сутками сидит в интернете? Вот включенное в безопасный режим расширение, которое блюрит мат в соцсетях.
- Создание модели, которая мало весит, но хорошо умеет детоксифицировать текст, позволит использовать её в любых подобных проектах

Команда, участники, роли

- Вероника Ганеева
- Яна Лабенская

- Поскольку нас всего две, мы постараемся распределять задачи равномерно и многим будем заниматься совместно – как вопросами организации и презентации проекта, так и изучением литературы и созданием решений.
- Поэтому наши задачи будут распределяться гибко, чтобы в случае какого-нибудь неприятного происшествия (например, как прямо сейчас) с одним из участников второй не оказался в тупике и без всего необходимого

Данные

- Так как мы берём соревнование с диалога, данные нам тоже предоставляются оттуда
- Это набор комментариев (3,539 штук в трейне с 1-3 детоксифицированными вариантами, 800 в тесте), и они достаточно короткие
- Как выглядят комментарии: содержат ошибки и опечатки, содержат редкие слова, содержат придуманные и сконструированные авторами оскорбительные слова
- К каждому комментарию в трейне нам предоставляется до трёх “детоксифицированных” версий

Бейзлайны

- **Бейзлайн один**: исключение слов по словарю
- Почему оно бейзлайн: просто, быстро, не нужно 16 гб видеокарты, обширный словарь, но не очень эффективно

- **Бейзлайн два**:

  t5, seq-to-seq, дообученная модель
- Почему оно бейзлайн: большая предобученная модель, сразу высокое качество, сложно побить (но требует вычислительных ресурсов)

Метрики оценки

- ***Точность передачи стиля*** (**STA**) оценивается с помощью классификатора на основе BERT (доработанного от Conversational Rubert), обученного на слиянии наборов токсичных комментариев на русском языке, собранных с 2ch.hk и с ok.ru.

  Зачем она нужна: комментарии должны потерять свою токсичность и мат, т.е. значительно изменить свой стиль

- ***Оценка сохранения значения*** (**SIM**) оценивается как косинусное сходство эмбеддингов предложений в LaBSE.

  Зачем она нужна: необходимо, чтобы комментарий без оскорблений передавал смысл изначального комментария. То есть текст о книге, которая автору комментария не понравилась, должен быть сведён к чему-то вроде “а мне эта книга не понравилась”.

Метрики оценки

- ***Оценка беглости*** **(FL)** оценивается с помощью классификатора беглости. Это модель на основе BERT, обученная отличать настоящие тексты, созданные пользователями, от искаженных. Для каждой пары предложений вычисляется вероятность искажения полученного и целевого предложений. Общая оценка беглости — это разница между этими двумя вероятностями: модель детоксикации должна выдавать текст, не уступающий по беглости исходному сообщению.

  Зачем она нам нужна: нежелательно, чтобы полученные в результате обработки предложения перестали выглядеть нормальным человеческим текстом, нужно стремиться к тому, чтобы они таким текстом выглядели. Обрабатывать следует максимально натурально, и эта метрика оценивает “натуральность” получившегося

Метрики оценки

- ***Совместная оценка***: три показателя, чтобы получить одно число, по которому можно сравнивать модели. Он рассчитывается как усредненное произведение STA, SIM и FL на уровне предложения: J = (STA * SIM * FL). Эта метрика будет использоваться для ранжирования моделей во время автоматической оценки.

  Зачем она нужна: нужно учитывать все три метрики для каждого варианта, чтобы понимать, какой вариант лучше. Эта метрика именно это и делает

План работ

1. **Реализовать препроцессинг**
    a. Очистка текста от нестандартных символов (эмодзи и т.п.)
    b. Очистка текста от пунктуации и пунктуационных эмодзи (:) и т.п); токенизация
    c. Экспериментально: реализация спеллчекинга

2. **Провести эксперименты с разными архитектурами и вариациями t5 (mt5, BYt5, etc.)**
    a. использовать предобученные модели из репозитория авторов и провести оценку
    b. дообучить модели на более подходящих для нас источниках и провести оценку

3. Поставить несколько экспериментов с обучением нейронных сетей:

- сочетание LSTM и CRF: хорошо показало себя в нашем прошлом проекте в задаче присваивания меток – здесь мы хотим попробовать использовать его для определения того, нужно ли нам в этом контексте (а контекст важен, иначе можно было бы взять словарь) удалять это слово

- RNN и их модификации: не имеют такого успеха на текстах, как LSTM, но возможно, нам не нужен будет широкий контекст для маркирования слов

4. Выбрать наш лучший результат и применить к нему подход transfer-learning для получения более компактной модели

Литература

1. Оригинальная статья о t5
2. Усовершенствованная архитектура t5, созданная авторами оригинальной уже после написания статьи: t5 1.1
3. mt5: t5, но мультиязыковой (предобучен на 101 языке, включая русский)
4. Альтернативный seq2seq метод text-editing: мы не перестраиваем текст с нуля, а пользуемся операциями вырезания/вставки/изменения для каждого токена

Литература

5. BYt5: m5, но вместо токенов на вход поставляются байты в UTF-8, а глубина кодировщика и декодировщика не выровнены, а специально разбалансированы: кодировщик имеет в три раза большую глубину
6. Архитектура C-LSTM для определения тональности слова по контексту
7. Метод усовершенствования предобученных моделей вида т5

Литература

8. Модель, хорошо дистиллирующая t5 в более миниатюрный вариант