ТУСУР
TUSUR UNIVERSITY

Оценка токсичности комментариев по шкале 1-10

Выполнил: Боднарчук Алексей Андреевич
(гр. 743-1)
Руководитель: Е.Ю. Костюченко, доцент
каф. КИБЭВС

Цель и задачи

Цель: Разработка модели, предсказывающей уровень токсичности.
Задачи:

- Изучить архитектуру RuBERT и методы регрессии
- Подготовить и очистить датасет
- Обучить модель
- Оценить результаты по метрикам MAE и RMSE
- Проанализировать ошибки

Актуальность

- Рост агрессии в онлайн-коммуникации
- Необходимость автоматической модерации
- Эффективность нейросетей и трансформеров
- Возможность количественной оценки токсичности

Среда и инструменты

- Язык: Python
- Среда: Google Colaboratory
- Библиотеки: PyTorch, Transformers, pandas, sklearn
- Датасет: Comments

Архитектура модели

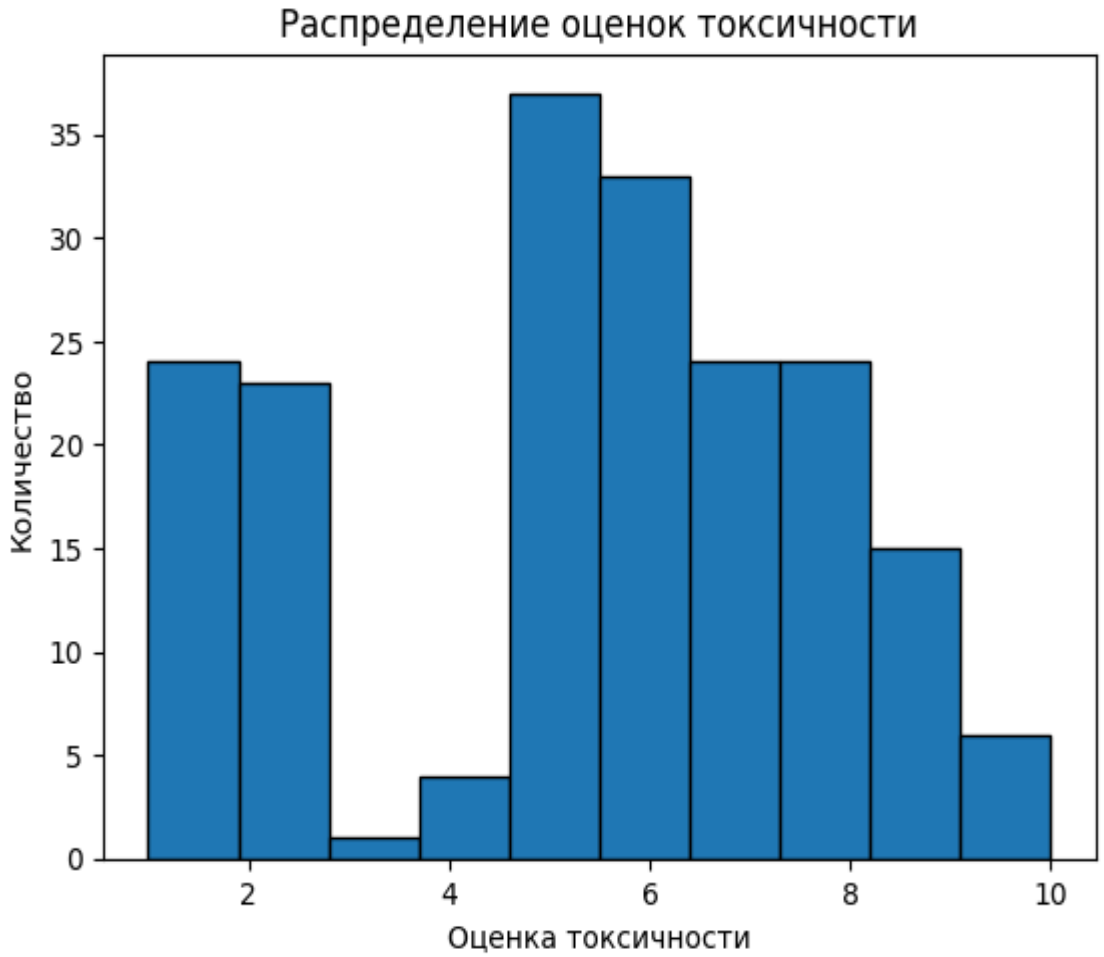
Используется предобученная модель RuBERT
Добавлен линейный регрессионный слой
На вход — токенизированный текст, на выход — число от 1 до 10

Архитектура модели:

```
RuBertRegressor(
  (bert): BertModel(
    (embeddings): BertEmbeddings(
      (word_embeddings): Embedding(120138, 768, padding_idx=0)
      (position_embeddings): Embedding(512, 768)
      (token_type_embeddings): Embedding(2, 768)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
    (encoder): BertEncoder(
      (layer): ModuleList(
        (0-11): 12 x BertLayer(
          (attention): BertAttention(
            (self): BertSdpaSelfAttention(
              (query): Linear(in_features=768, out_features=768, bias=True)
              (key): Linear(in_features=768, out_features=768, bias=True)
              (value): Linear(in_features=768, out_features=768, bias=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
            (output): BertSelfOutput(
              (dense): Linear(in_features=768, out_features=768, bias=True)
              (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
          )
          (intermediate): BertIntermediate(
            (dense): Linear(in_features=768, out_features=3072, bias=True)
            (intermediate_act_fn): GELUActivation()
          )
          (output): BertOutput(
            (dense): Linear(in_features=3072, out_features=768, bias=True)
            (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
      )
    )
    (pooler): BertPooler(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (activation): Tanh()
    )
  )
  (dropout): Dropout(p=0.3, inplace=False)
  (out): Linear(in_features=768, out_features=1, bias=True)
)
```

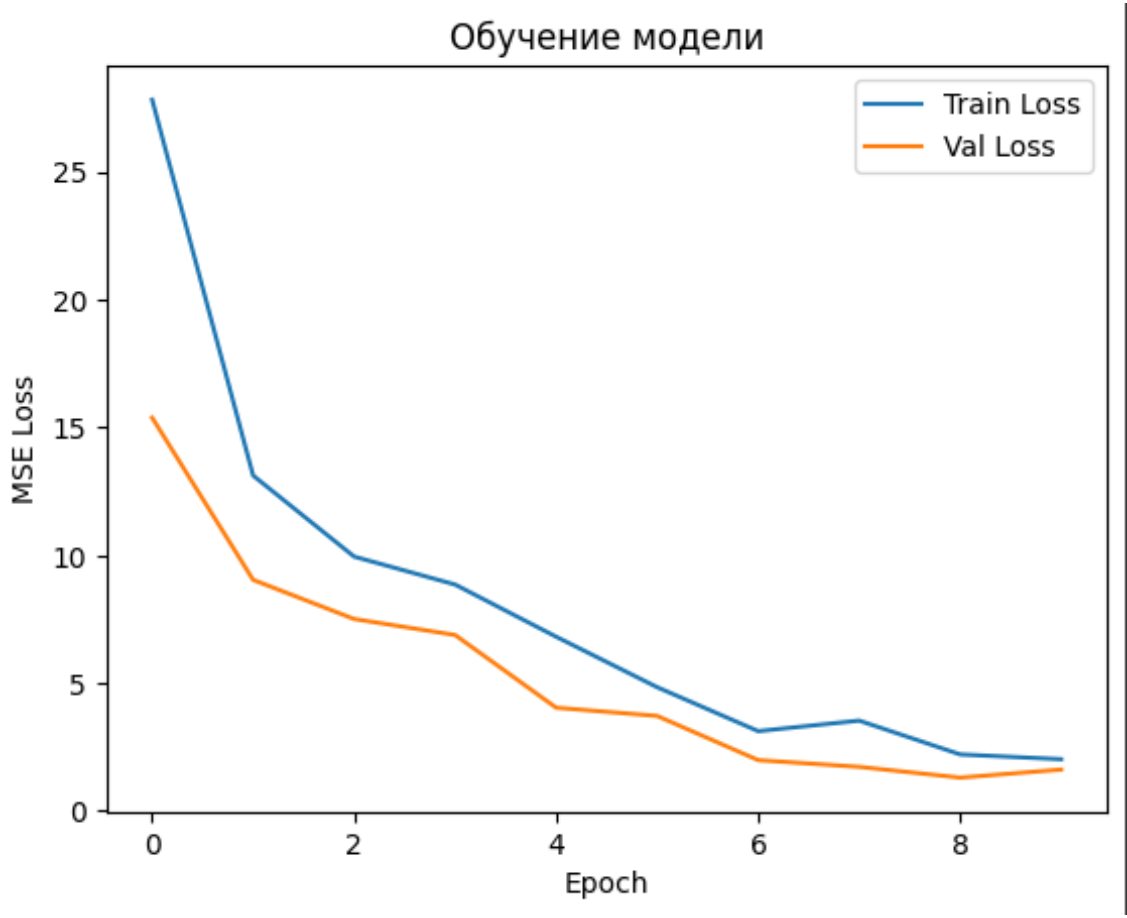
Подготовка данных

- Очистка текста: удаление лишних символов
- Токенизация с AutoTokenizer
- Разделение данных на train/val
- Балансировка классов



Результаты обучения

- MAE: 1.104
- RMSE: 1.252
- Модель обучалась стабильно, без переобучения



Примеры предсказаний

ТУСУР
TUSUR UNIVERSITY

- Полученные на примерах оценки токсичности:

```
==================================================
Оценка комментариев вашей моделью
==================================================

Комментарий: "Это худшее видео, что я видел в жизни!"
Ваша оценка токсичности (1-10): 3.0399999618530273

Комментарий: "Спасибо за интересный ролик!"
Ваша оценка токсичности (1-10): 0.9599999785423279

Комментарий: "Ты полный идиот, не позорься!"
Ваша оценка токсичности (1-10): 7.0

Комментарий: "Что за убогое объяснение темы"
Ваша оценка токсичности (1-10): 6.71999979019165

Комментарий: "Научись нормально разговаривать"
Ваша оценка токсичности (1-10): 6.880000114440918
```

Выводы

- Модель успешно решает задачу оценки токсичности
- Предсказания соответствуют смыслу текста
- Средняя ошибка менее 1.2 баллов
- Перспективы: дообучение, увеличение датасета, анализ сарказма

ТУСУР
TUSUR UNIVERSITY

СПАСИБО
ЗА ВНИМАНИЕ!

tusur.ru