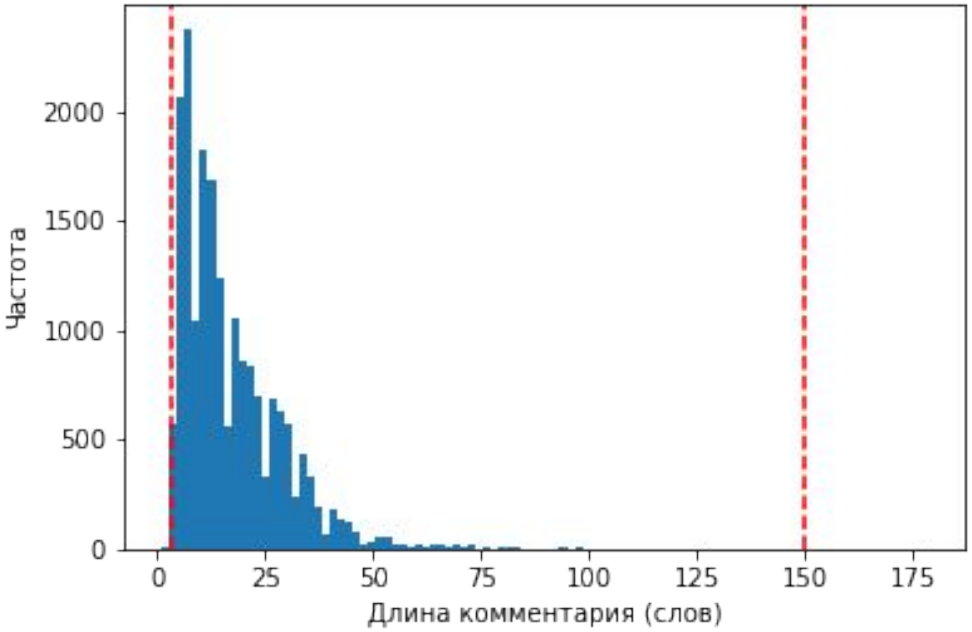
Amogus

HSE Data Science Hack

EDA. Работа с текстом.

Настройка стоп - слов под датасет

EDA. Работа с выбросами



Препроцессинг

- Используем разные датасеты для предсказания сентимента и категории отзыва
- Для сентимента удаляем дубликаты, отзывы с неопределенной категорией, а также отзывы, у которых для одного отзыва несколько сентиментов.
  После всех преобразований получилось 5943 наблюдений
- Для категорий удаляем дубликаты, отзывы с неопределенной категорией.
  Преобразуем две категории в столбец с одной категорией
  После всех преобразований получилось 8662 наблюдений

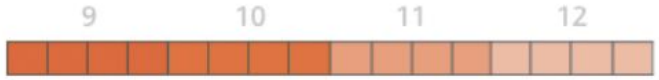
Общая идея

- Предобработка комментариев
  - Работа с дубликатами
- Baseline: CatBoost
  - Различные вариации CatBoost + Optuna и эмбендингов
- Основные модели: предобученный Bert
  - Раздельные модели для sentiment и category, а также совместная модель для двух таргетов.
  - Композиция моделей

Предсказание категории

- LaBSE с двумя головами - для мультикласс категории и сентимента.
- Поля категория1 и категория2 преобразовывались в вектор с индикаторами принадлежности каждому из классов.
- Для дубликатов с разными таргетами вектора объединялись
- 5 эпох файнтюнинга, lr=5e-5
- Hidden state с последних 4 слоев энкодера в качестве признаков для

Concat Last
Four Hidden



Логи обучения

Метрики ROC - AUC

Модель	**Sentiment**	**Category**
CatBoost	0.9	0.56
CatBoost + TFiDF sentiment	0.63	0.57
CatBoost word2vec	0.88	0.65
Bert	**0.97**	0.82
Bert with 2 losses	**0.94**	**0.87**

Спасибо за внимание!