

HSE DATASCIENCE HACK, 2023

# Прогнозирование этичности банков

команда NAMM

# Что мы сделали в начале

## Шаг 1

Провели  
препроцессинг  
данных

Стеминг, токенизация и  
своя небольшая Bag-of-  
words модель.  
NLTK

## Шаг 2

Построили  
нейронную сеть  
на PyTorch для  
первой задачи

3 слоя, обучалась 5  
секунд

## Шаг 3

Проверили  
качество первой  
модели на  
валидации

Получили roc-auc 0.93

## Шаг 4

Повторили  
предыдущие шаги для  
второй задачи

Разделили на 4 класса

## Шаг 5

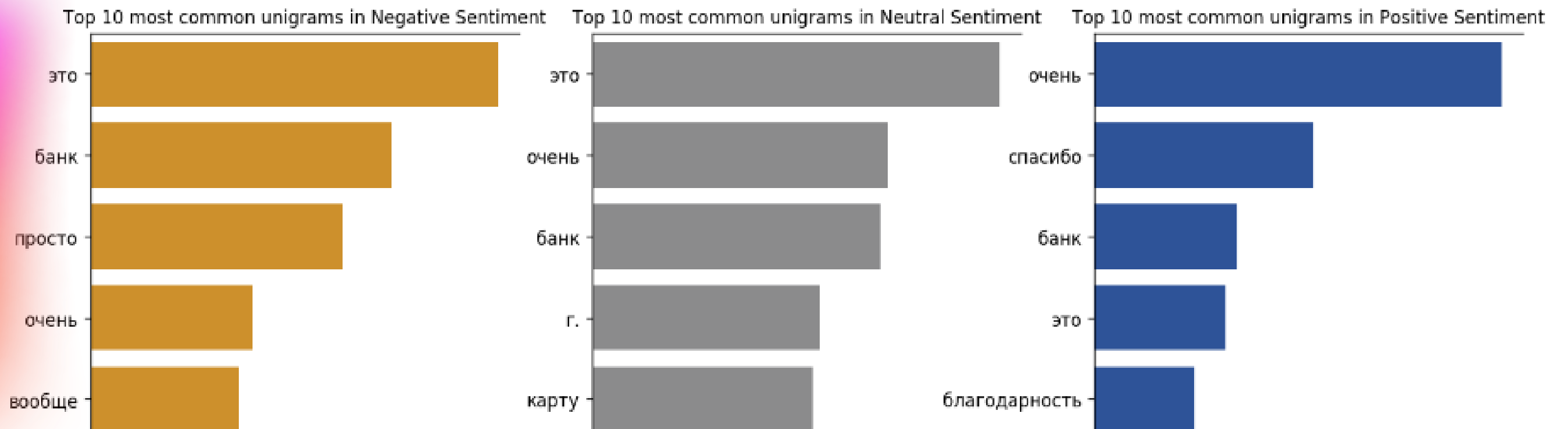
Проверили  
качество второй  
модели на  
валидации

Получили roc-auc 0.81

На часах 17:00, что делать  
дальше...

# Провели EDA-анализ

- Уникальных предложений: 7k
- Уникальных записей: 13k
- 51% / 34% / 15% - Negative / Positive / Neutral



# Попробовали ruRoBERTa от Сбера

BERT, обученный на большом корпусе текста, только на задачу восстановления маскированных токенов, на большом батч-сайзе и с токенизатором BBPE от ruGPT-2.

## Сопоставимое качество

Получили схожее качество на первой задаче на более сложной модели. Как следствие, было решено вернуться к более простой и быстрообучаемой модели

# Попробовали ruBERT-sentiment от Deep Pavlov

1. Использовался RuBERT для Sentiment Analysis от DeepPavlov
2. Модель дообучалась с заменной головой
3. Оба таргета сразу (при обучении - сумма лоссов по двум задачам)
4. Результаты: 0,8 на Задаче 1 и 0,45 на задаче 2
5. Модель оказалась очень тяжелой, более того, плохо обучаемой на совместной задаче. Было решено от нее отказаться

# Невоплощенные идеи

- BPE + TF-IDF + CatBoost: использование GB в классическом подходе с Byte-Pair токенизацией
- Дообучение Bert-like модели на данных из banki.csv: в том же формате, что и в оригинальной статье, для лучшей работы с лексикой в комментариях
- Использование данных из banki.csv при обучении мультизадачной модели
- Предсказать на основе наших 2 моделей данные для banki.csv и изучить корреляцию таргета с оценкой



# В главных ролях:



Максим Егоров

так себе шутник



Никита Курдюков

недопонятый гений



Сергей Кушнерюк

злодей британец



Арслан Шахназаров

мужик с зарплатой до колен