Классификация эмоций в текстах

Team 39

Команда

Участник	Роль	Функции
Кубракова Екатерина	Team Lead, Data Scientist	общая координация проекта, EDA, подготовка данных, тестирование базовых архитектур
Дяминова Эльвира	Data Scientist, ML Engineer	подбор гиперпараметров, обучение модели классификатора эмоций на базе T5, подготовка презентации
Лилиом Елизавета	Data Scientist, NLP Engineer	аугментация данных, тестирование разных подходов к расширению базового набора данных, балансировка классов, тестирование моделей Paraphrasing, оптимизация baseline решения
Карнюшин Виталий	MLOps	развертывание модели, подготовка инфраструктуры для инференса, сравнительный анализ инференс-серверов, оформление репозитория

Основные характеристики задачи

Задача	Классификация эмоций в текстах
Язык текстов	Русский
Тип задачи	Multi-class + Multi-label
Тип модели	LM (до 1 млрд параметров)
Оцениваемая метрика	F1-score (weighted)
Количество классов	7 (anger, disgust, fear, joy, sadness, surprise, neutral)
Домен данных	Транскрибированные тексты (ASR)
Особенности домена	Отсутствие пунктуации и капитализации, ошибки транскрибации
Стек технологий	pytorch, transformers, huggingface

Оригинальный датасет

Датасет состоит из

train: 43410
validation: 5426
test: 8742

Было замечено, что предоставленный датасет является переводом англоязычного датасета на 28 эмоций – [GoEmotions]

Поэтому часть русскоязычных предложения выглядят своеобразно:

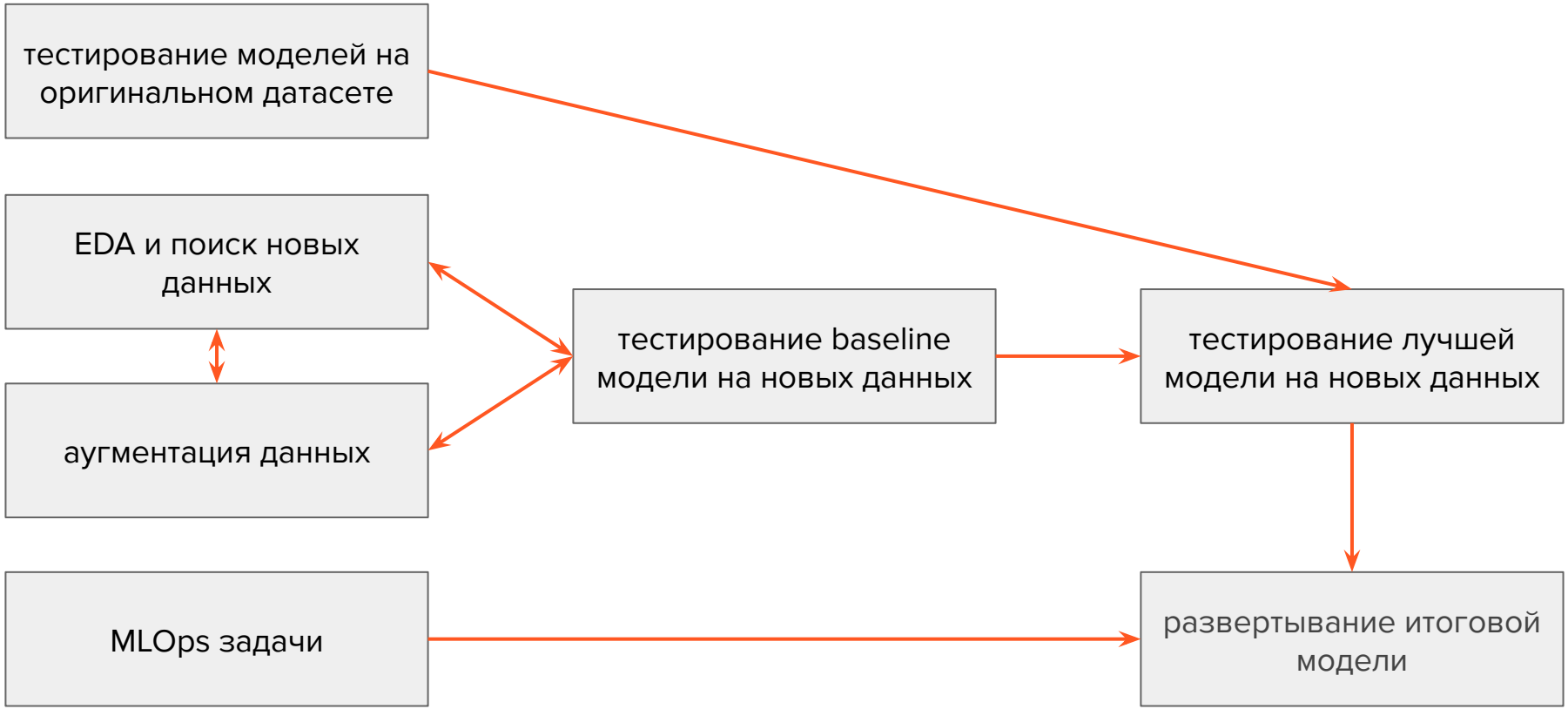
- *Dirty Southern Wankers —> Грязные южные дрочники*
- *WHY THE FUCK IS BAYLESS ISOING —> КАКОГО НАХРАНА БЭЙЛЕССКАЯ ИЗОИНЦИЯ?*
- *Sack, shaft, and tip. The trifecta. —> Мешок, стержень и наконечник. Трифекта.*
- и т.д.

Датасет на русском изначально имеет ряд погрешностей, поэтому наша основная цель – приблизиться к 0.6 на тестовых данных.

Сравнение переводов

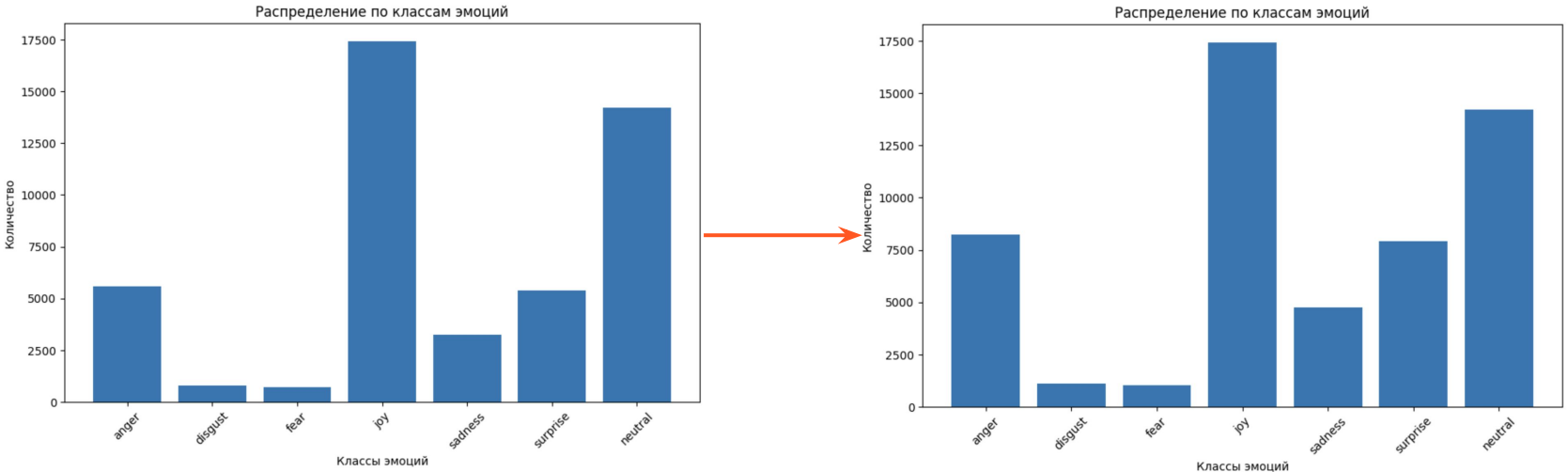
Текст на английском GoEmotions	Перевод от Криптонита	Перевод от Helsinki-NLP
My favourite food is anything I didn't have to cook myself.	Моя любимая еда — это все, что мне не приходилось готовить самому.	Моя любимая еда - это то, что мне не нужно было готовить самому.
Now if he does off himself, everyone will think hes having a laugh screwing with people instead of actually dead	Теперь, если он покончит с собой, все будут думать, что он смеется, трахая людей, а не на самом деле мертв.	А теперь, если он сойдет с ума, все подумают, что он смеется над людьми вместо того, чтобы умереть.
WHY THE FUCK IS BAYLESS ISOING	КАКОГО НАХРАНА БЭЙЛЕССКАЯ ИЗОИНЦИЯ?	Почему **censored** хреново изображается?
To make her feel threatened	Чтобы она почувствовала угрозу	Чтобы она чувствовала себя под угрозой.
Dirty Southern Wankers	Грязные южные дрочники	Грязные южные петухи

RoadMap проекта



Подготовка данных

- *Exploratory Data Analysis* (EDA). Лемматизация и токенизация данных, составление облака слов для оценки ключевых и наиболее частотных слов. Оценка распределения слов по классам.
- Аугментация данных за счет двойного перевода *«ru – en – ru»*. Использовались модели от *Helsinki-NLP/opus-mt-...-...*
  - *fear & disgust:* 43410 -> 44707
  - *anger & fear & disgust & sadness & surprise*: 43410 -> 50745



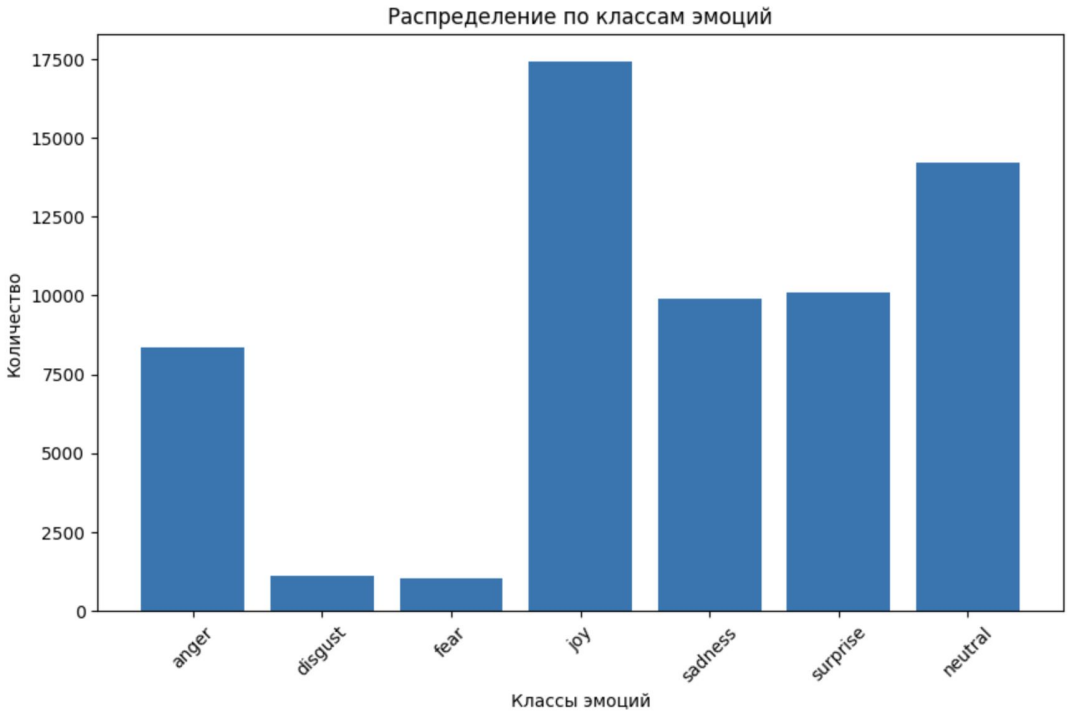
Подготовка данных

- Поиск новых датасетов и их апробация на baseline модели. В итоге был взят датасет с твитами на английском языке:

  sadness: 5165
  surprise: 2187
  anger: 110

  Для перевода использовалась модель от *Helsinki-NLP*.

  Это позволило увеличить тренировочный датасет до 58204.



Модели от ai-forever

Модели от ai-forever	Num Parameters	Training Data Volume	Tokenizer
ruBert-base	178 M	30 GB	BPE
ruBert-large	427 M	30 GB	BPE
ruRoberta-large	355 M	250 GB	BBPE
ruT5-base	222 M	300 GB	BPE
ruT5-large	737 M	300 GB	BPE

Результаты text classification на тесте

Подход	f1 weighted	Датасет
ruBert-base (baseline)	0.54681	оригинальный набор данных
ruBert-large	0.55736	оригинальный набор данных
ruBert-base	0.55010	аугментация *fear&disgust*
ruBert-large	0.55195	аугментация *fear&disgust*
ruBert-large	0.56865	аугментация 5 классов «*ru–en–ru*»
ruRoberta-large	0.59800	аугментация 5 классов «*ru–en–ru*»
ruRoberta-large	0.59814	аугментация 5 классов «ru–en–ru» + доп. данные от твитов

text-to-text А что, если...?

«черт кажется я случайно купил боксерский поединок с оплатой за просмотр»

anger	disgust	fear	joy	sadness	surprise	neutral
1	0	0	0	1	0	0

[1, 0, 0, 0, 1, 0, 0] → «гнев, грусть»

Результаты на тесте:

Подход	f1 weighted	Датасет
ruT5-base	0.54899	оригинальный набор данных
ruT5-large	0.54509	оригинальный набор данных

Итоговый подход

model:

- ruRoberta-large
- epochs = 4

tokenizer:

- max_length = 32
- batch_size = 64

```
Epoch: 0
Train loss: 0.2750432781088891
Valid loss: 0.23896921617002342
Epoch: 1
Train loss: 0.23207176006843697
Valid loss: 0.2328041517854336
Epoch: 2
Train loss: 0.21175837721769233
Valid loss: 0.24288928890647601
Epoch: 3
Train loss: 0.18939521847919902
Valid loss: 0.23543907814289458
```

Результаты на тесте:

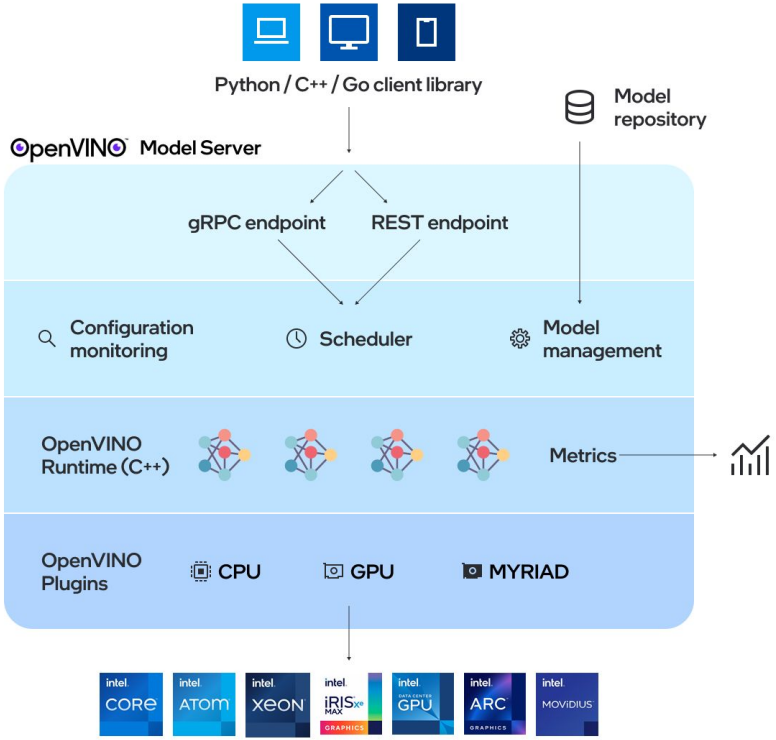
Подход	f1 weighted	Датасет
ai-forever/ruRoberta-large	0.59814	аугментация 5 классов «*ru–en–ru*» + доп. данные от твитов

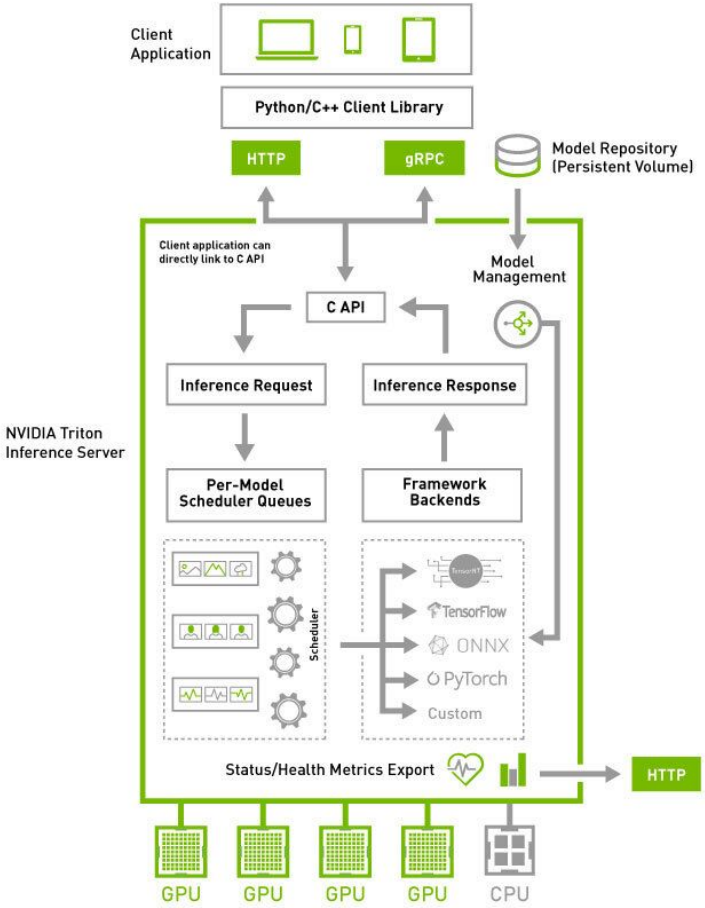
Infer like a Pro

Для задачи развертывания выбрали для сравнения два популярных инференс-сервера:

Характеристика	OpenVINO Model Server	Triton Inference Server
Производитель	Intel	NVIDIA
Основное назначение	Оптимизирован для инференса на CPU и Intel-ускорителях.	Оптимизирован для инференса на GPU и поддерживает широкий спектр аппаратуры.
Целевая аудитория	Пользователи Intel-хардвера (например, CPU, VPU, FPGA).	Пользователи GPU NVIDIA, но также поддерживаются CPU.
Форматы моделей	OpenVINO IR, ONNX, TensorFlow, PyTorch, PaddlePaddle, JAX/Flax	TensorRT, ONNX, TensorFlow, PyTorch, OpenVINO, Python, DALI, FIL, vLLM
Оптимизация производительности	Использует OpenVINO Toolkit для ускорения инференса на Intel-хардвере.	Поддерживает оптимизацию моделей через TensorRT для GPU.
Масштабирование	Масштабируется с использованием Kubernetes и Docker.	Есть встроенные механизмы масштабирования и оптимизации распределения нагрузки.

Infer like a Pro





Спасибо за внимание