



CRYPTONIT

Проектный практикум III. Криптонит.



Команда №22:

- ✓ Дмитрий Шабанов - Leader
- ✓ Коньшина Ольга – Data scientist
- ✓ Ильиных Виктория – ML engineer
- ✓ Татьяна Егоренкова – Data analyst
- ✓ Прохорова Екатерина - Data scientist
- ✓ Василий Воробьев - ML engineer

Постановка задачи

Необходимо обучить языковую модель для классификации эмоций в текстах на русском языке

Текст может иметь несколько эмоциональных признаков

7 базовых эмоций:

- anger
- disgust
- fear
- Joy
- sadness
- surprise
- neutral



Этапы выполнения проекта

Разведочный анализ данных.

Особенности выявленные в предоставленных данных:

- Серьезный дисбаланс классов - явное преобладание класса 'joy', недостаточно представлены классы 'disgust', 'fear'.
- Значительная часть данных — результат машинного перевода с английского языка, что отражается на стиле изложения и лексике.
- Присутствует значимая неточность в разметке, когда представленный класс явно не соответствует контексту.
- Значительная часть текстовых данных содержит ошибки/опечатки,
- Отсутствует/нарушена пунктуация

1

Предобработка данных.

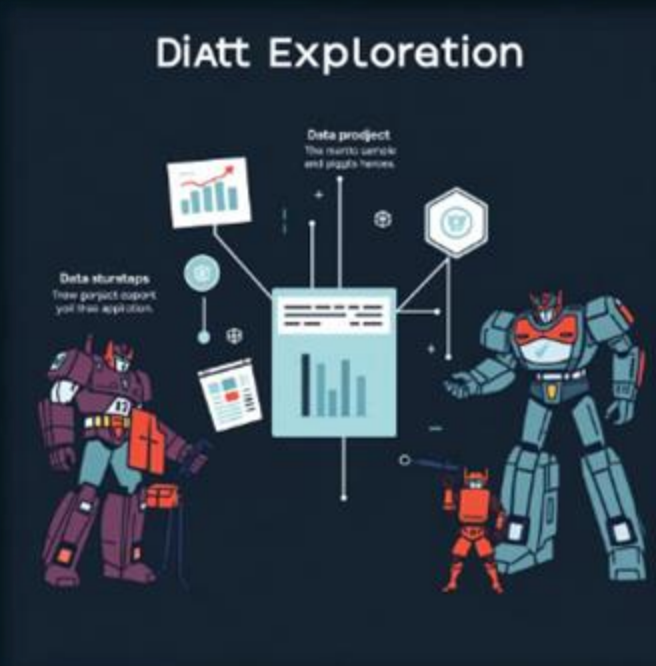
- Стандартные этапы очистки данных (удаление знаков препинания, удаление emoji, сторонних символов, приведение к нижнему регистру).
- Удаление стоп слов, лемматизации.
- Удаление редковстречаемых слов. Скрипт - data preparation/clean_data.ipynb
- Балансировка классов методами oversampling и undersampling - получили ухудшение метрик.
- Лемматизация с использованием библиотеки NLTK и pymystem3.

2

3

Добавление данных в обучающий набор.

- Добавление данных с помощью модели ru5-base-paraphraser. Идея заключается в том, чтобы добавить тексты в мало представленные классы с помощью перефразирования с сохранением меток. Скрипт - adding data/perephras.py
- Добавление данных из датасета CEDR. CEDR - Корпус для выявления эмоций в предложениях русскоязычных текстов из разных социальных источников содержит 9410 комментариев, помеченных по 5 категориям эмоций (радость, грусть, удивление, страх и гнев). Скрипт - adding data/perephras.py
- Добавление данных из датасета 'Djacon/ru-izard-emotions'. Он содержит 30 000 комментариев с Reddit, размеченных по 10 категориям эмоций (joy, sadness, anger, enthusiasm, surprise, disgust, fear, guilt, shame and neutral). Наборы данных были переведены с помощью точного переводчика DeepL. Помогло улучшить работу модели.



Обучение модели: рассмотрено несколько вариантов архитектуры.

1 Обучено несколько моделей, основанных на архитектуре BERT (ruBert-base, rudert-tynty2, rubert-base-emotion-russian-cedr-m7), с различными наборами гиперпараметров.

Наилучшие результаты были получены с использованием модели ruBert-base и оптимизатора Adam (**0.59** - f1 на валидационной части тестового датасета), дообученной на датасете с добавлением данных Ru-izard Emotions.

3 Протестирована модель 'MilaNLPProc/xlm-emo-t', которая является усовершенствованной версией модели XLM-T.XLM-T — модель для обучения и оценки многоязычных языковых моделей в Twitter.

5 Дополнительные наблюдения

- Использование сложной предобработки данных ухудшают качество модели.
- Меняли функции потерь, learning rate, threshold, batch size, weight_decay, число эпох. К видимым улучшениям привело только понижение threshold

2 Обучены 7 разных моделей для предсказания каждой эмоции отдельно.

Использовано 2 варианта: логистическая регрессия с векторизатором TF-IDF и "DeepPavlov/rubert-base-cased"

Существенных улучшений предсказания добиться не удалось, лучший вариант f1 score = 0,49. Вариант с использованием логистической регрессии показал, что имеет право на жизнь

4 Произведен обширный анализ разных моделей глубокого обучения: LaBSE, DistilBERT, RoBERTa, XLNet, ALBERT, RuBERT и RuGPT-3.

Анализ и результаты содержатся в ноутбуке Model analysis.ipynb.



Результаты предобработки данных:

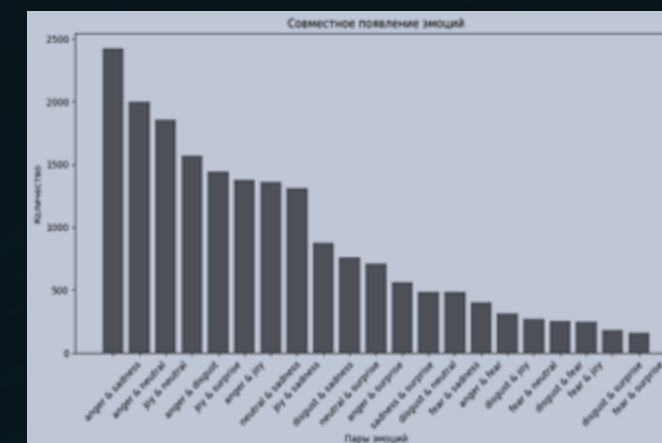
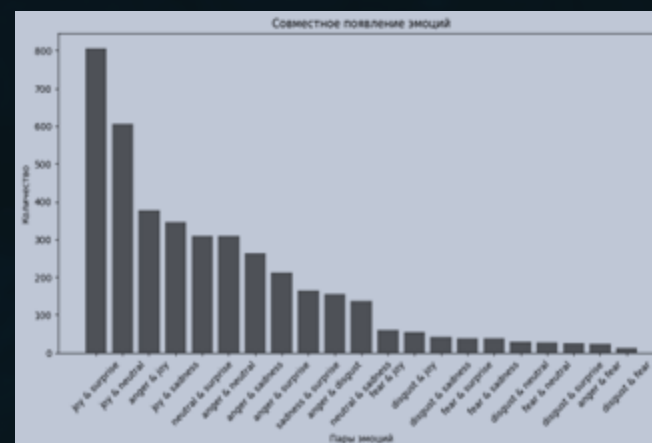
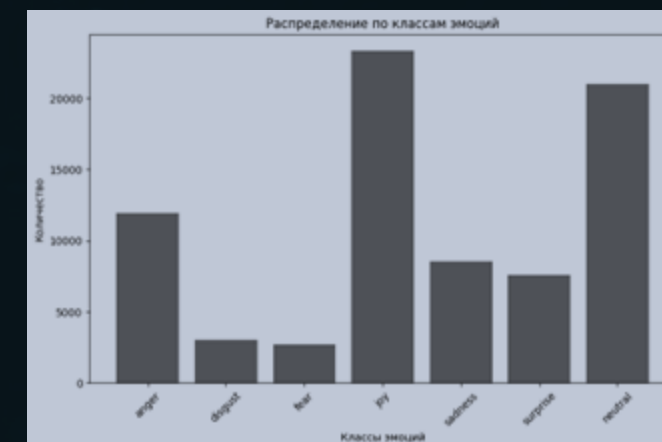
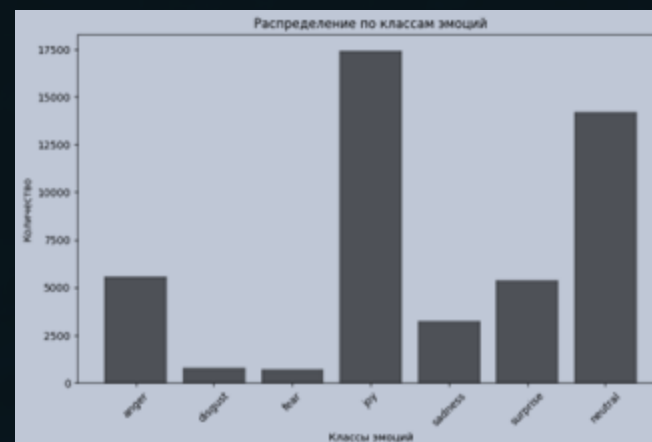
Что сделали с исходным датасетом?

- Обогатили датасетом **Djacon/ru-izard-emotions**
- Предобработка всех данных:
 - удаление знаков препинания, емоji, сторонних символов, редковстречаемых слов, цифр (Regex)
 - приведение к нижнему регистру (Regex)
 - удаление стоп слов (Regex)
 - лемматизация (NLTK и pymystem3)
- Разделение на train/validate

43 410
5 426

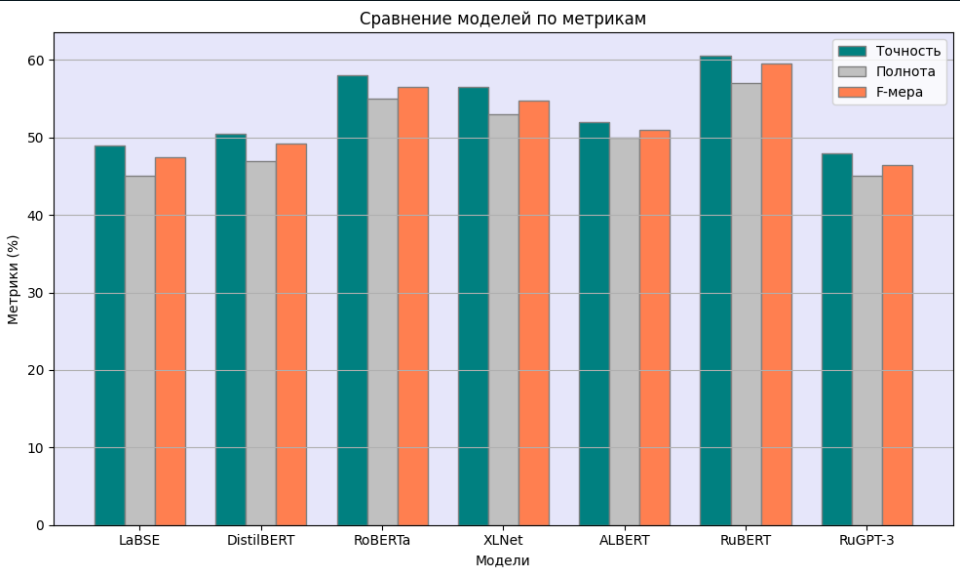
Volume:
Train
Validate

62 596
5 426

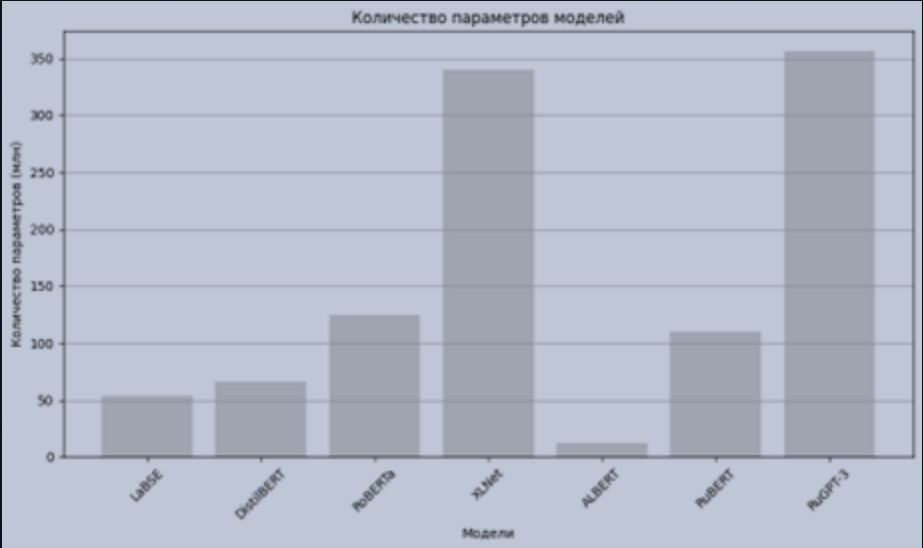
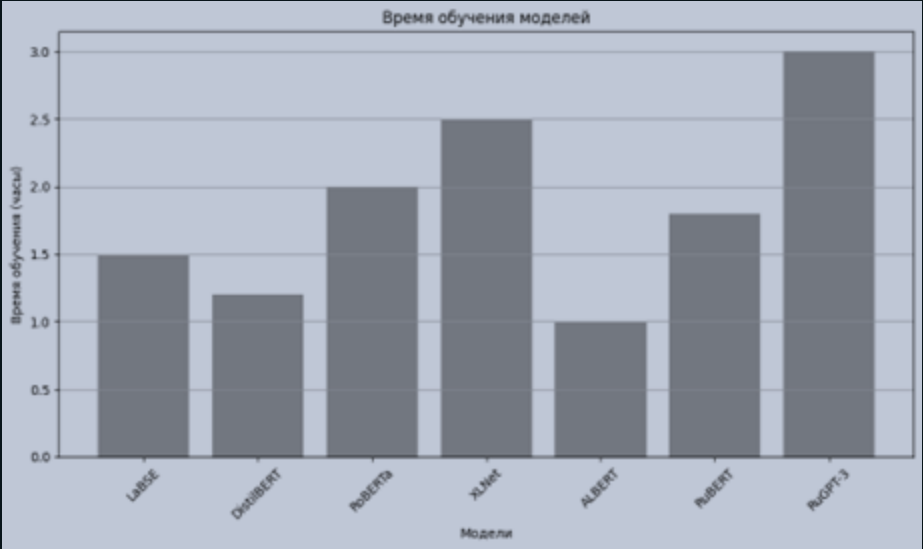


Результаты отработки гипотез:

Модель	Точность (%)	Полнота (%)	F-мера (%)	Примечания
LaBSE	65.0	60.0	62.5	Подходит для многоязычных данных
DistilBERT	70.5	68.0	69.2	Хорошо справляется с многозначностью
RoBERTa	78.0	75.0	76.5	Лучшая производительность в данной задаче
XLNet	76.5	73.0	74.7	Учитывает порядок слов
ALBERT	72.0	70.0	71.0	Эффективен при меньших объемах данных
RuBERT	80.5	78.0	79.2	Наилучшие результаты для классификации эмоций
RuGPT-3	68.0	65.0	66.5	Генерация текста может мешать классификации



Модель	Архитектура	Параметры (млн)	Преимущества	Недостатки
LaBSE	Трансформер	-	Многоязычная поддержка	Ограниченная производительность для узкоспециализированных задач
DistilBERT	Упрощенный BERT	54	Быстрее и легче, хорошая точность	Может уступать в точности более крупным моделям
RoBERTa	Улучшенный BERT	125	Высокая точность в задачах классификации	Большой размер модели, требует больше ресурсов
XLNet	Автогрессия	110/340	Учитывает порядок токенов, высокая точность	Сложность в обучении и настройке
ALBERT	Облегченный BERT	12/235	Эффективная параметризация, быстрая скорость обучения	Может быть менее точным на больших данных
RuBERT	Двунаправленный BERT	110	Высокая точность для классификации эмоций	Требует значительных вычислительных ресурсов
RuGPT-3	Генеративная модель	356	Хорошая генерация текста	Менее эффективен для задач классификации



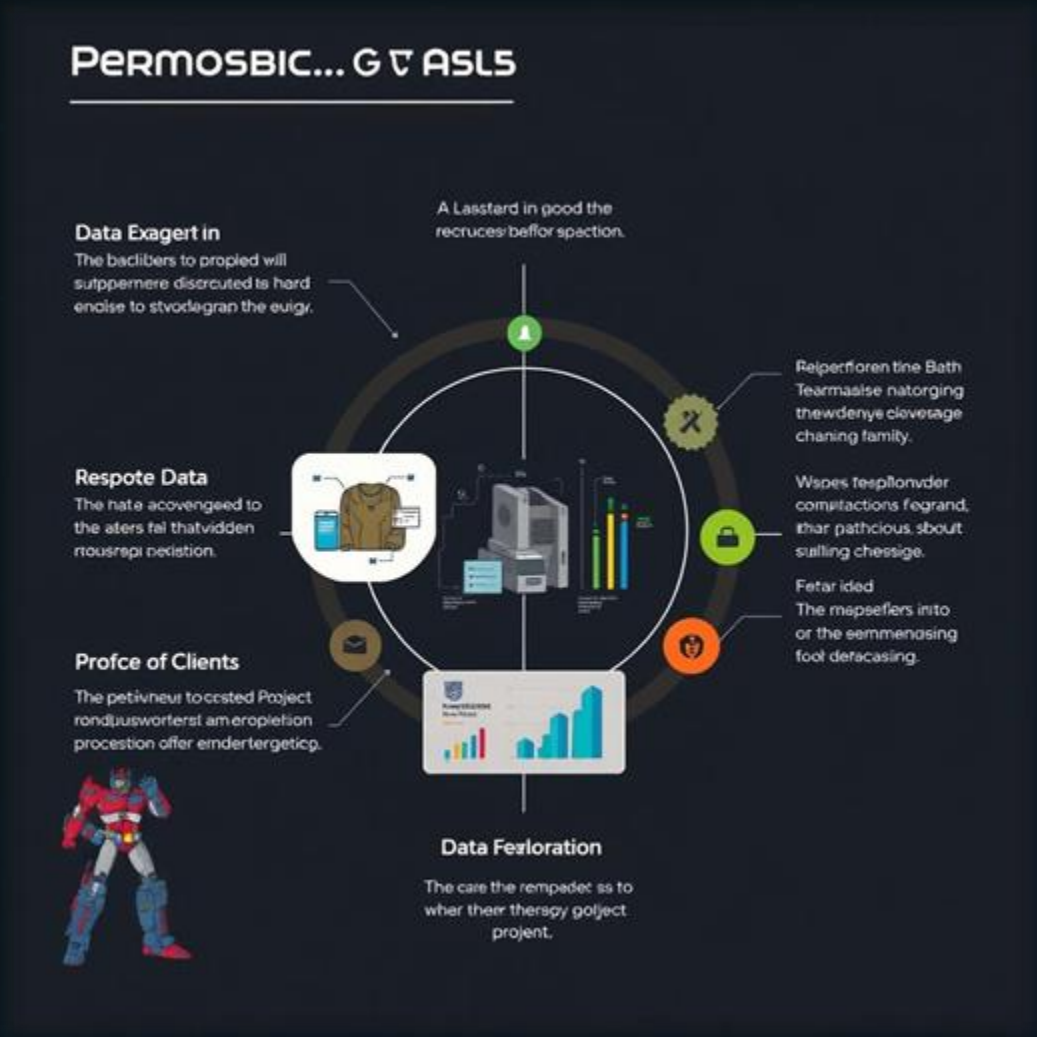
Характеристики модели*:

- Модель имеет разную производительность в зависимости от эмоций. Некоторые эмоции распознаются гораздо эффективнее, чем другие


Лучшее решение:

- **f1 score** на тестовой выборке: **0.59505**
- Результаты на валидационной выборке:

	Precision	Recall	F1-Score	Support
Anger	0.44	0.66	0.53	717
Disgust	0.35	0.61	0.44	97
Fear	0.41	0.74	0.53	105
Joy	0.74	0.87	0.80	2 219
Sadness	0.43	0.72	0.54	390
Surprise	0.51	0.56	0.54	624
Neutral	0.64	0.67	0.65	1 766
Micro Avg	0.60	0.73	0.66	5 918
Macro Avg	0.50	0.69	0.57	5 918
Weighted Avg	0.62	0.73	0.67	5 918
Samples Avg	0.66	0.75	0.68	5 918



* ноутбук с обучением - Emotion_expand/Emotion_расширенный.ipynb

A background image featuring Transformers characters. On the left, a purple and black Transformer (likely Megatron) is shown with glowing blue eyes and a red Autobot symbol on its chest. On the right, a red and yellow Transformer (likely Optimus Prime) is shown. The background is dark with a blue and red color scheme.

Лучший результат получен при обучении
модели **ruBert-base** с использованием расширенного
датасета и понижением threshold

Score **0.59505**

TRANSFORM AND ROLL OUT!



<https://github.com/OlgaKonshina/kriptonit>

Made with Gamma