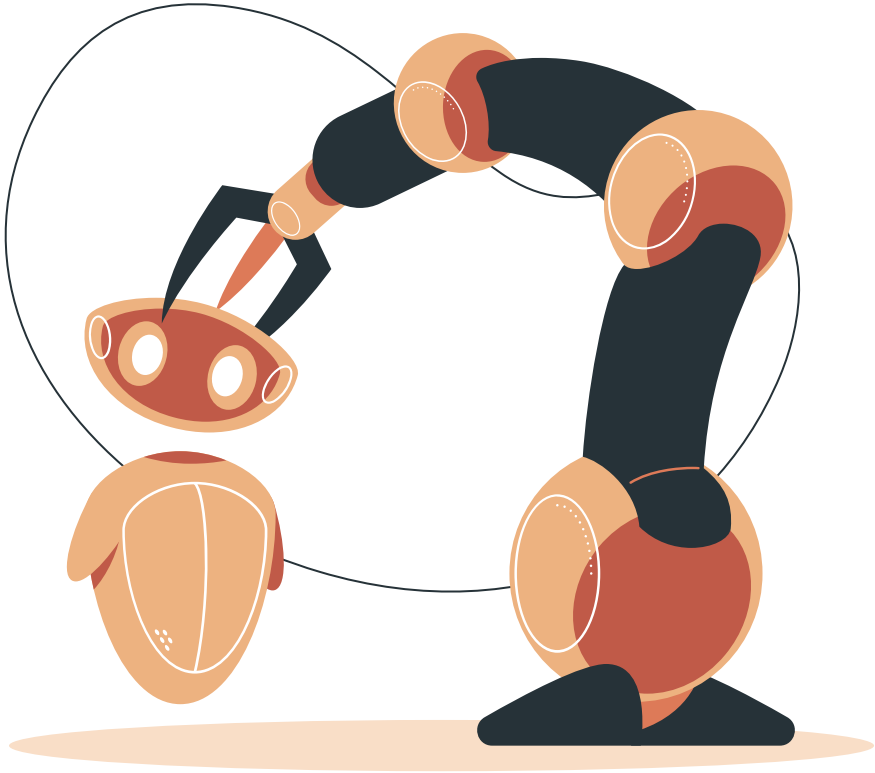
Сравнение эффективности **базовых моделей** и **моделей, предобученных для решения задачи анализа тональности текста**, при дообучении для анализа тональности именованных сущностей

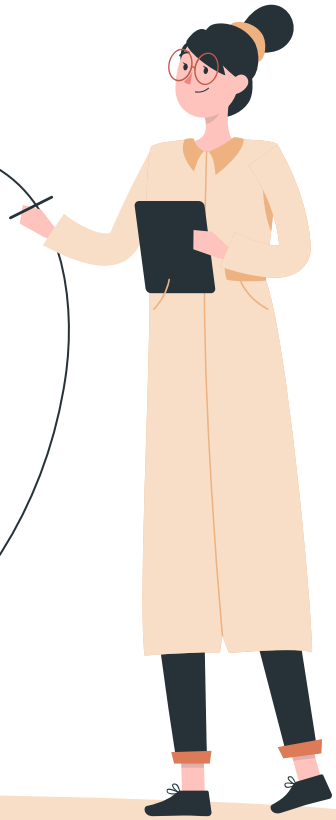
Кравчук Мария
Ожогова Элина
Тыщишина Таисия



1

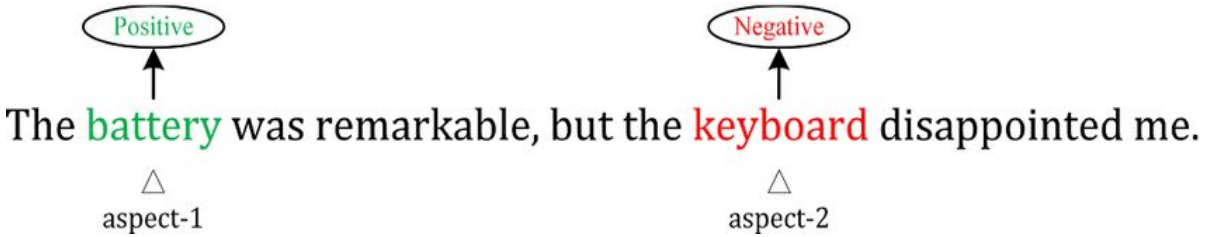
Повторение

- Постановка задачи

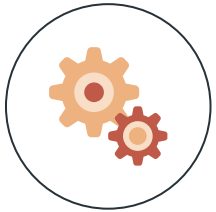


TSC

Таргетированный анализ тональности (TSC, *target-dependent sentiment classification*) – это подзадача анализа тональности, нацеленная на определение отношения к конкретным **сущностям** и их **свойствам** или **темам**.



Предыдущие подходы



Классическое ML

- тщательное конструирование признаков
- составление словарей эмоционально окрашенной лексики
- **F1 = 63.3**

Эмбеддинги и DL

- разработка нейронных архитектур
- тонкая настройка базовых языковых моделей
- **F1 = 75.8**

Особенности новостных текстов

1

Нейтральный стиль

Язык новостных статей зачастую нейтрален, авторы не выражают своё отношение эксплицитно

2

Разные интерпретации

Разные читатели могут по-разному оценивать отношение статьи к целевой сущности

[Hamborg et al., 2021]

2

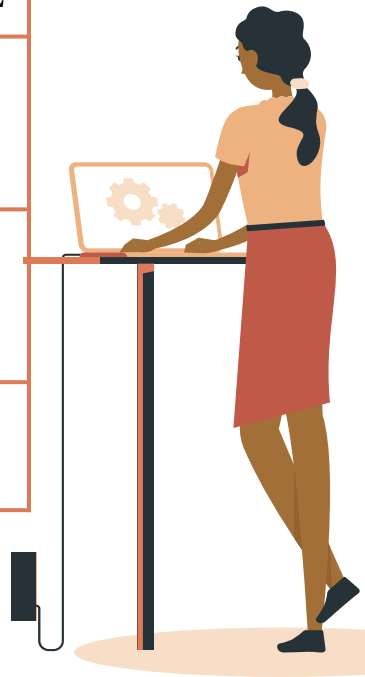
Датасет

- RuSentNE2023

О датасете

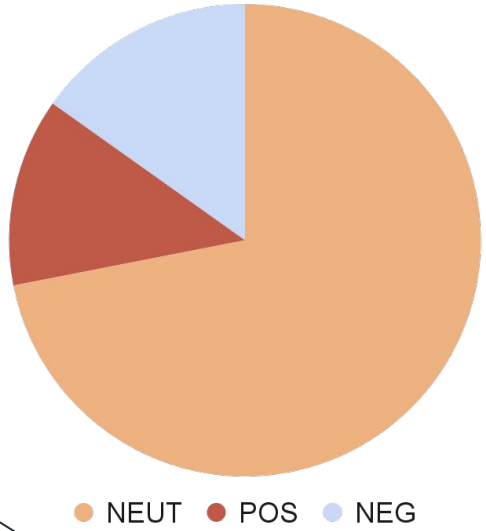
- Итоговая разметка: **сущность**, **тип сущности**, **метка класса** (-1, 0, 1)

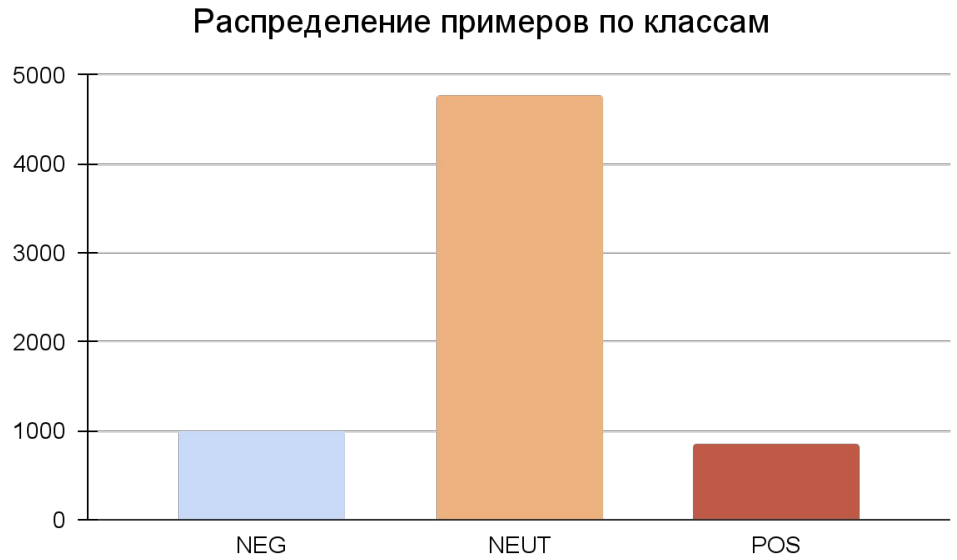
SENTENCE	ENTITY	ENTITY_TAG	LABEL
Восемь бадминтонисток были дисквалифицированы на Олимпийских играх	бадминтонисток	PROFESSION	-1
Ещё недавно, после завершения матча сборной России и Португалии, Юрий приезжал в Тамбов с семьёй.	Португалии	COUNTRY	0
Владислав первым заметил возгорание и начал тушить его.	Владислав	PERSON	1



Распределение классов в датасете

- Большинство примеров в обучающей выборке относятся к нейтральному классу





Расширение датасета

- Для борьбы с дисбалансом классов был расширен набор данных положительного и отрицательного классов с помощью автоматического перефразирования.
- Модель для перефразирования: rut5-base-paraphraser* (парафразер для предложений на русском языке, обученный на корпусах субтитров и новостных заголовков).

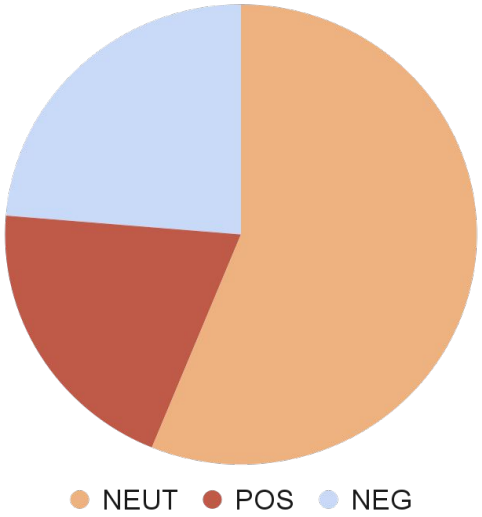
```
print(paraphrase('Владислав первым заметил возгорание и начал тушить его.'))
```

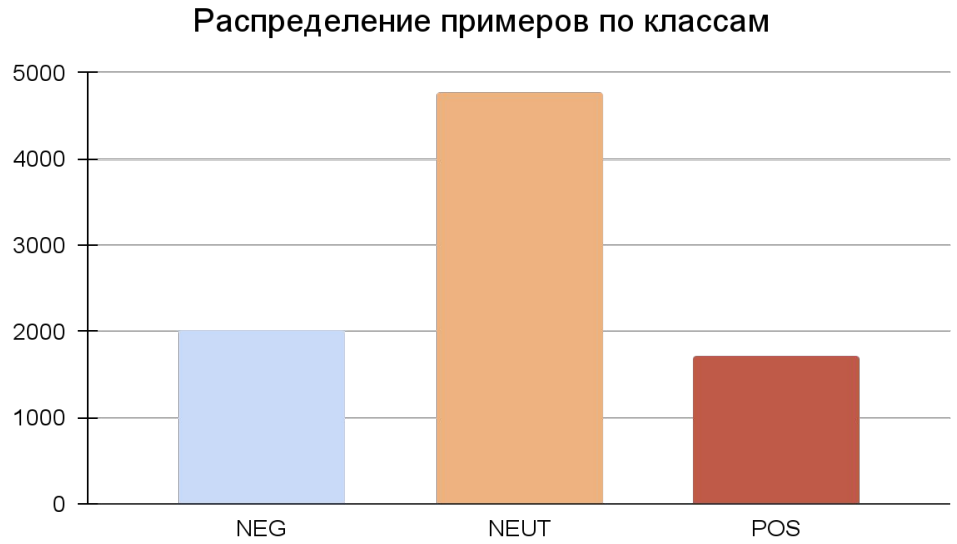
```
Владислав первый заметил пожар и начал его тушить.
```

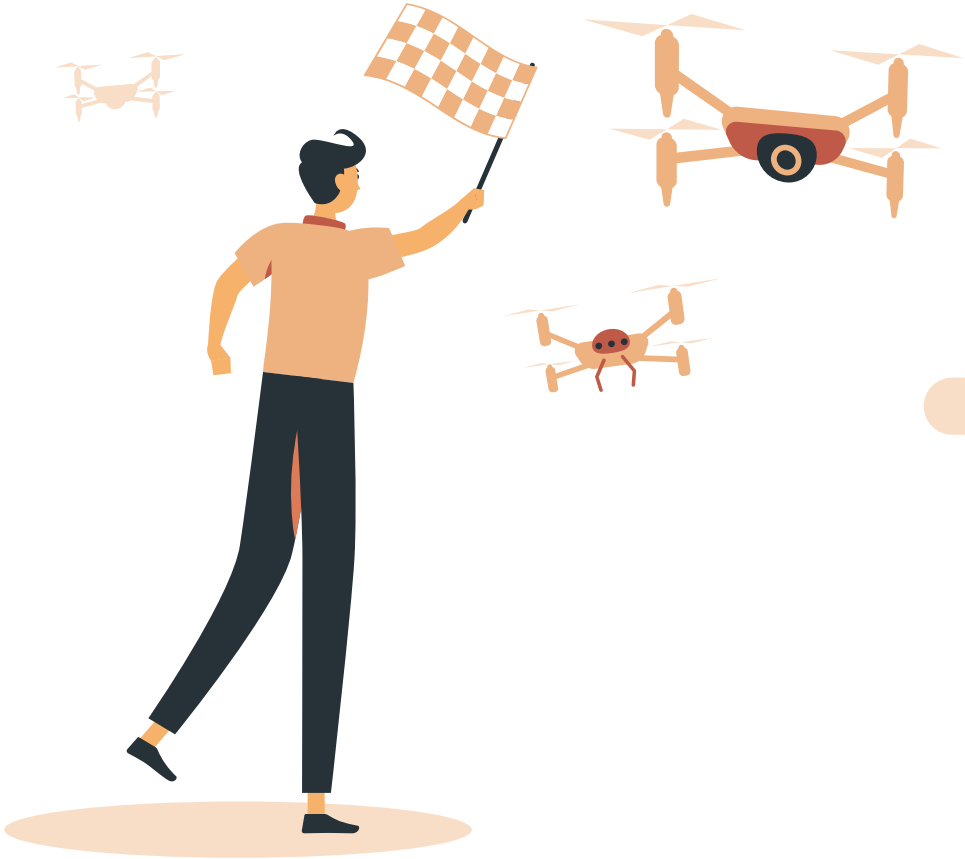
* https://huggingface.co/cointegrated/rut5-base-paraphraser

Распределение классов в датасете

- В связи с более равномерным распределением примеров по классам ожидается повышение качества при дообучении на обновленном датасете







Наше решение

3

Цель

- Сравнить качество базовых и предобученных (анализу тональности текстов) моделей при дообучении анализу тональности к именованным сущностям

Задачи

1. Выравнивание количества данных в разных классах с помощью автоматического перефразирования
2. Дообучение базовой LLM на полученном датасете
3. Дообучение предобученной анализу тональности текстов LLM
4. Ансамбль моделей

Гипотеза

- Ожидается повышение качества при дообучении на расширенном датасете.
- Модели, предобученные на анализ тональности, будут справляться с анализом тональности к именованным сущностям лучше базовых.

Базовая модель

- В качестве базовой непредобученной модели была выбрана: DeepPavlov/rubert-base-cased*.
  - Модель RuBERT (12-layer, 768-hidden, 12-heads, 180M параметров) была обучена на русскоязычной Википедии и новостных текстах.

* https://huggingface.co/DeepPavlov/rubert-base-cased

Предобученные модели

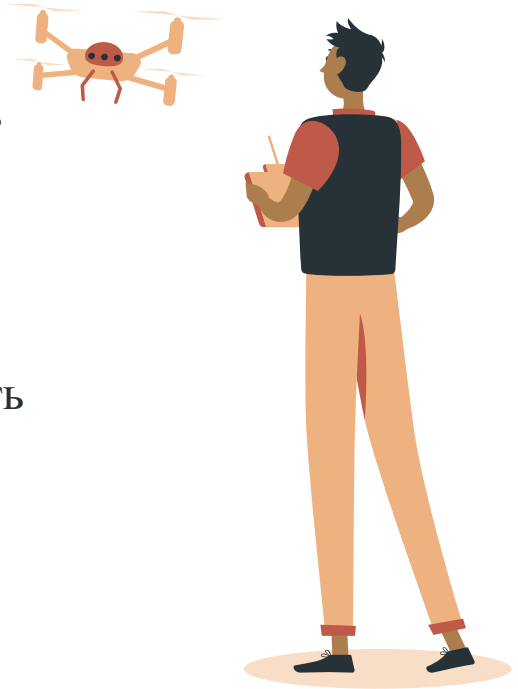
Все предобученные модели были настроены на задачу распознавания тональности в русскоязычных текстах (классификация по трем классам):

- blanchefort/rubert-base-cased-sentiment*
- seara/rubert-base-cased-russian-sentiment**
- r1char9/rubert-base-cased-russian-sentiment***
- cointegrated/rubert-tiny-sentiment-balanced****

* https://huggingface.co/blanchefort/rubert-base-cased-sentiment
** https://huggingface.co/seara/rubert-base-cased-russian-sentiment
*** https://huggingface.co/r1char9/rubert-base-cased-russian-sentiment
**** https://huggingface.co/cointegrated/rubert-tiny-sentiment-balanced

Дообучение

- Задача анализа тональности как классификация пары предложений.
- Были протестированы различные варианты вопросов, однако наилучший результат показало решение (Golubev et al. 2023):
  - На вход подаются два предложения, разделенные токеном [SEP]:
    - вопрос *“Как относятся к X?”* где X – сущность в дательном падеже;
    - текст предложения.



Промптинг

- В ходе работы были предложены различные промпты, однако качества выше, чем в [Golubev et al. 2023] добиться не удалось.
- Качество базовой модели на валидационной выборке в зависимости от промпта:

	F1(P,N,0)-macro	F1(P,N)-macro
Как относятся к X?	**0.69**	**0.45**
Что думают о X?	0.68	0.44
Каково мнение о X?	0.66	0.43
Как оценивают X?	**0.69**	**0.45**

Результаты

4

Результаты на валидации

- Для сравнения качества все модели были дообучены как на базовом датасете, так и на расширенном датасете

	base F1_PN0	enlarged F1_PN0	base F1_PN	enlarged F1_PN
Модель без предобучения	0.67	–	0.43	–
Предобученные модели:				
seara/rubert-base-cased-russian-sentiment	0.69	0.67	0.46	0.43
r1char9/rubert-base-cased-russian-sentiment	0.67	0.66	0.45	0.42
cointegrated/rubert-tiny-sentiment-balanced	0.54	0.55	0.33	0.39
blanchefort/rubert-base-cased-sentiment	0.36	0.28	0.11	0.00

Результаты на тестовой выборке

	F1(P,N,0)-macro	F1(P,N)-macro
Модель без предобучения	64.77	54.21

Предобученные модели:

blanchefort/rubert-base-cased-sentiment	36.41	14.05
seara/rubert-base-cased-russian-sentiment	**63.13**	**52.37**
r1char9/rubert-base-cased-russian-sentiment	32.82	8.97
cointegrated/rubert-tiny-sentiment-balanced	46.81	30.91

Лучшие результаты на CodaLab

- Макро F-мера по двум классам (**F1(P,N)-macro**) на тестовой выборке на платформе CodaLab

	Базовый датасет	Расширенный датасет
Модель без предобучения	**55.24**	**54.21**
Предобученная модель	**54.49**	**52.37**

Ансамбль моделей

- гипотеза: повышение качества

Модели в ансамбле

- Качество моделей на тестовой выборке на платформе CodaLab:

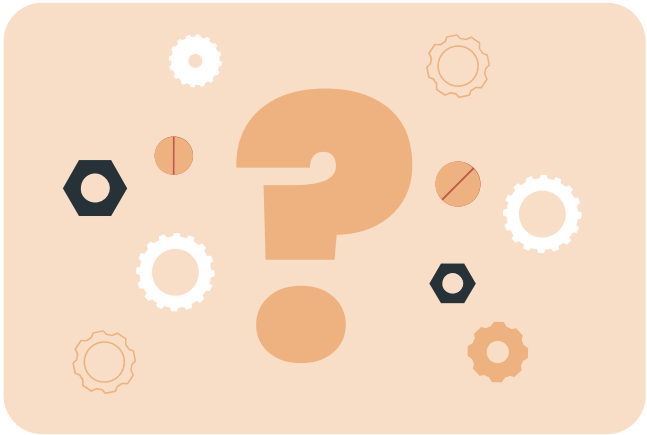
	F1(P,N,0)-macro	F1(P,N)-macro
Модель без предобучения	**62.19**	**50.62**
seara/rubert-base-cased-russian-sentiment	**61.87**	**49.77**
cointegrated/rubert-tiny-sentiment-balanced	**47.28**	**34.00**

Архитектура ансамбля: 1

Ансамбль из трёх моделей:

- базовая модель;
- seara/rubert-base-cased-russian-sentiment;
- cointegrated/rubert-tiny-sentiment-balanced.

Ансамбль выбирает тот класс, за который **проголосовало большинство** моделей.

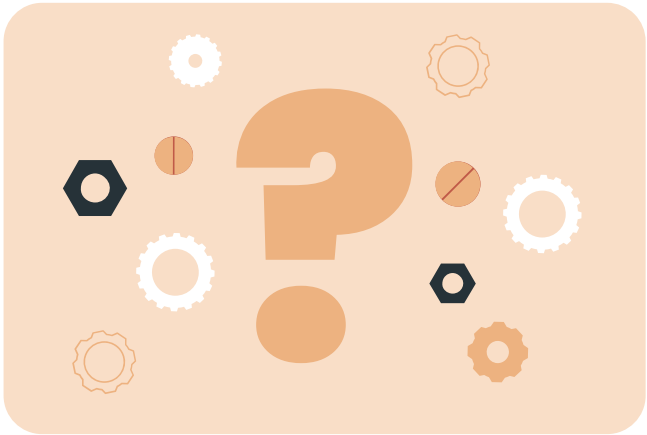


Результаты на CodaLab

	F1(P,N,0)-macro	F1(P,N)-macro
Ансамбль из трёх моделей	**62.34**	**50.73**
Ср. с моделями, вошедшими в ансамбль:		
Модель без предобучения	**62.19**	**50.62**
seara/rubert-base-cased-russian-sentiment	**61.87**	**49.77**
cointegrated/rubert-tiny-sentiment-balanced	**47.28**	**34.00**

Архитектура ансамбля: 2

- Модели из ансамбля предсказывают ответы на валидационной (или обучающей!) выборке.
- Далее линейная модель (логистическая регрессия) обучается на предсказанных ансамблем метках классов и реальных ответах.
- Веса классов сбалансированы в зависимости от размера класса.



Результаты на CodaLab: итог

	F1(P,N,0)-macro	F1(P,N)-macro
Второй ансамбль	**62.19**	**50.62**
Первый ансамбль	**62.34**	**50.73**
Модель без предобучения на базовом датасете: лучший результат	**65.33**	**55.24**

Обсуждение

- Использование моделей, предобученных для решения анализа тональности текста, не дало ожидаемого роста качества.
  Вероятно, несмотря на внешнюю схожесть, в сущности SA и TSC – разные задачи, и предобучение только путает модели. Стоит также отметить, что предобучение проводилось на текстах из отзывов и постов из социальных сетей, которые значительно отличаются по стилю от новостных текстов.

Обсуждение

- Расширение датасета не помогло решить проблему несбалансированности выборки. Вероятно, чтобы преодолеть порог F-меры в 64-65% для трёх классов (54-55% для двух классов), необходимо расширить датасет естественным образом (либо применить методы взвешивания классов, как было проделано в конкурсных решения на RuSentNE-2023).

Литература

- Golubev et al., 2023 – Golubev, A., Rusnachenko, N., & Loukachevitch, N. (2023). RuSentNE-2023: Evaluating entity-oriented sentiment analysis on Russian news texts. arXiv preprint arXiv:2305.17679.
- Hamborg et al., 2021 – Hamborg, F., Donnay, K., & Gipp, B. (2021). Towards target-dependent sentiment classification in news articles. In Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part II 16 (pp. 156-166). Springer International Publishing.