

Сравнение эффективности русскоязычных и
мультязычных моделей
при решении задачи **анализа тональности**
к именованным сущностям в **новостных**
текстах

Даниил Бурмистров
Елизавета Дорофеева
Виктория Яровая

Flashback #1 : проблемы

[Loukachevitch, Rusnachenko 2018]

- мнения высказаны разными субъектами ;
- один и тот же субъект назван по-разному (*Россия - Российская Федерация*) ;
- несколько именованных сущностей с разными тональностями ;
- преобладание нейтрального класса ;
- имплицитный характер тональности (чего не скажешь о корпусах типа *Laptop and Restaurant*) .

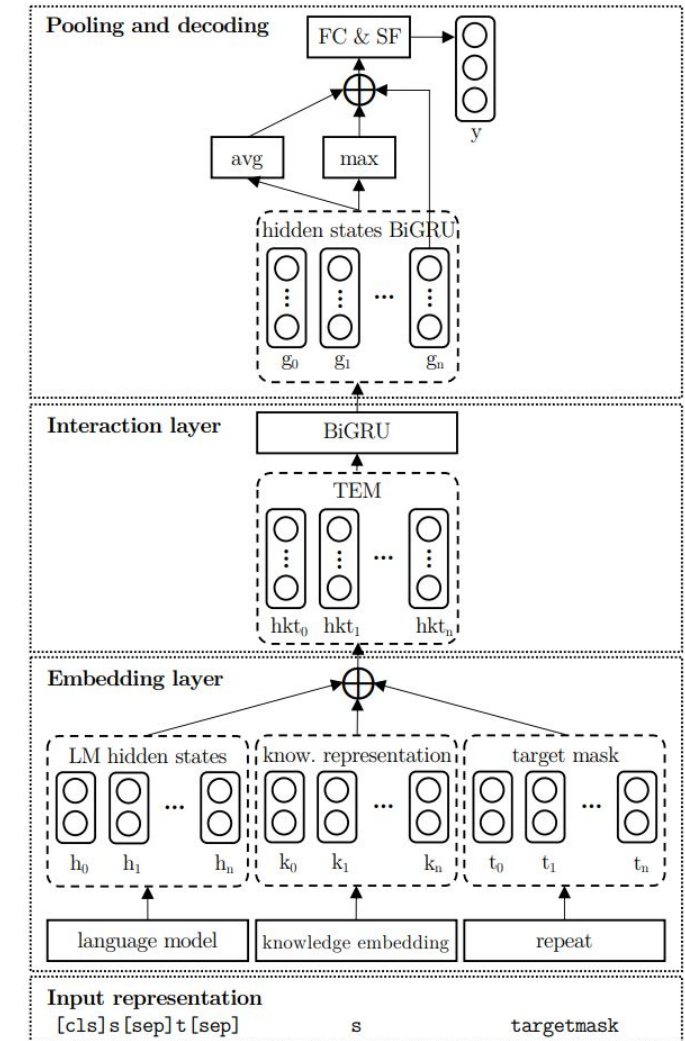
Flashback #2: предыдущие исследования

[Hamborg et al., 2021]:

“The vast majority of recently proposed TSC approaches employ **BERT** and focus on devising specialized **down-stream architectures**”;

Additional measures:

- **domain-specific** LM **finetuning** prior to the TSC finetuning;
- use of **external knowledge**, such as sentiment or emotion dictionaries;
- **rule-based** sentiment systems;
- use of **all mentions** of a target and/or related targets in a document;
- explicit **encoding of syntactic information**.



Flashback #2: предыдущие исследования

RUSENTNE-2023:

- *baseline*:
conversational **ru**BERT_{base} + задача = классификация пары предложений + промт = "Как относятся к X?";
- 1 место:
использование 2 моделей (½ специализированно русскоязычная): одна для исходного предложения, другая для предложения с масками + ф-ция потерь = взвешенная кросс-энтропия;
- 2 место:
RuRoBERTa_{large} + игры с выделением сущностей + метод взвешивания классов WINS.

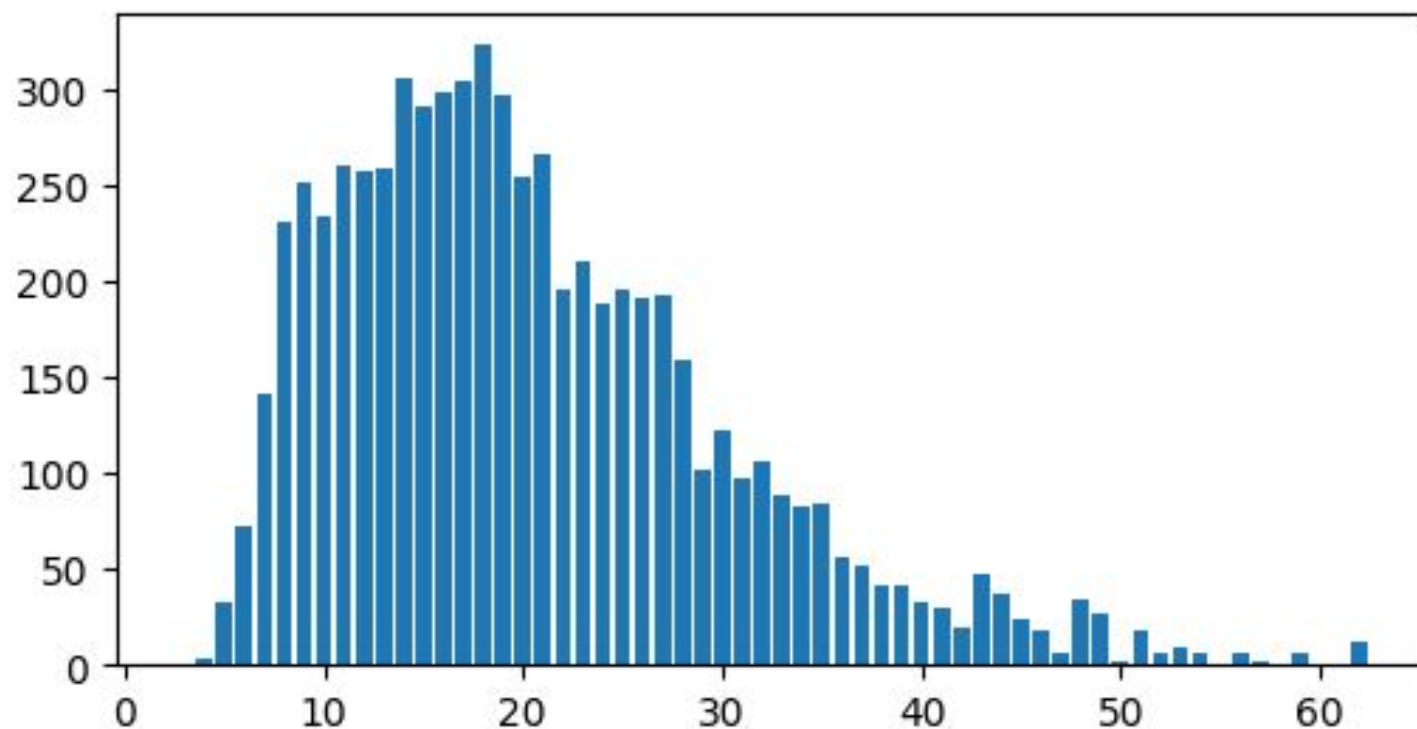
[Solomatina, Loukachevitch 2024]:
prompt-tuning + аугментация данных позитивного и негативного класса

Анализ данных RuSentNE2023

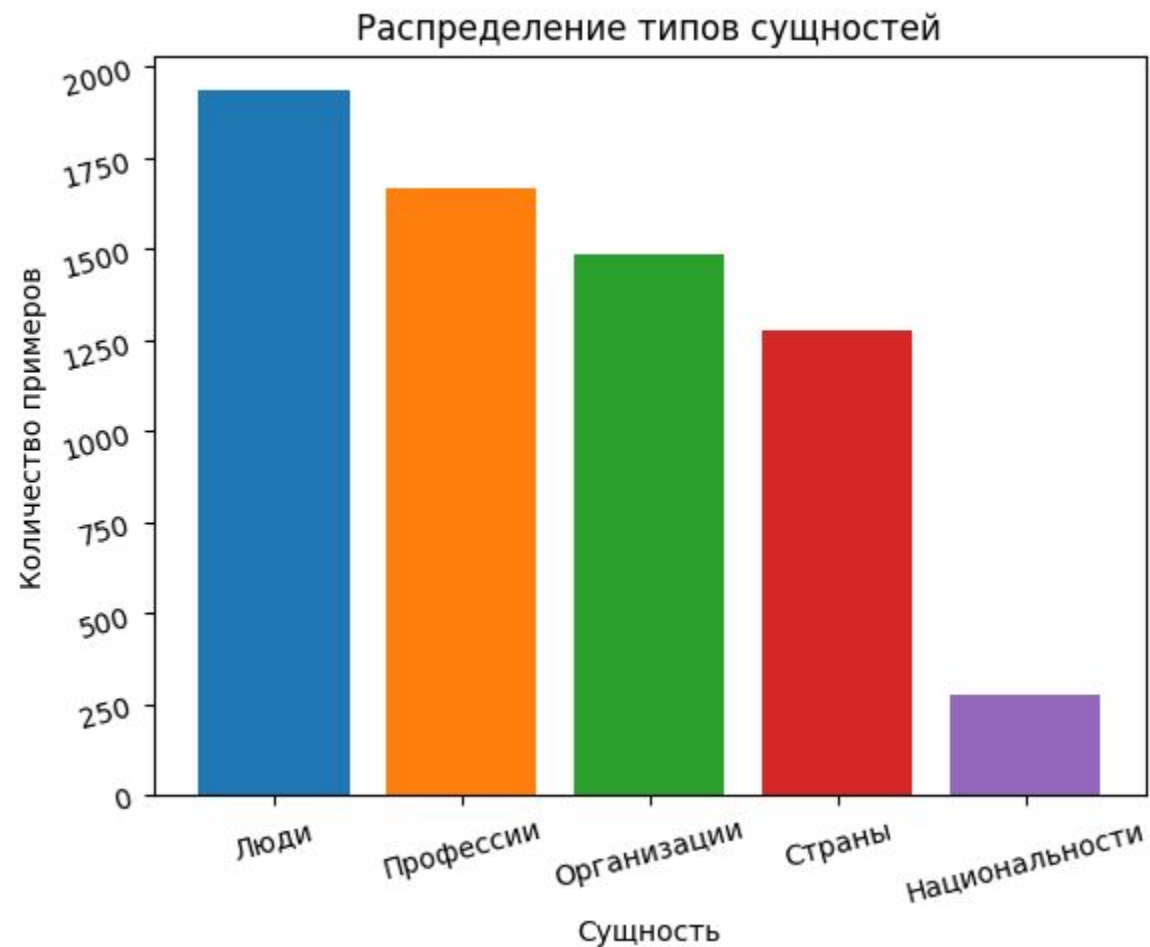
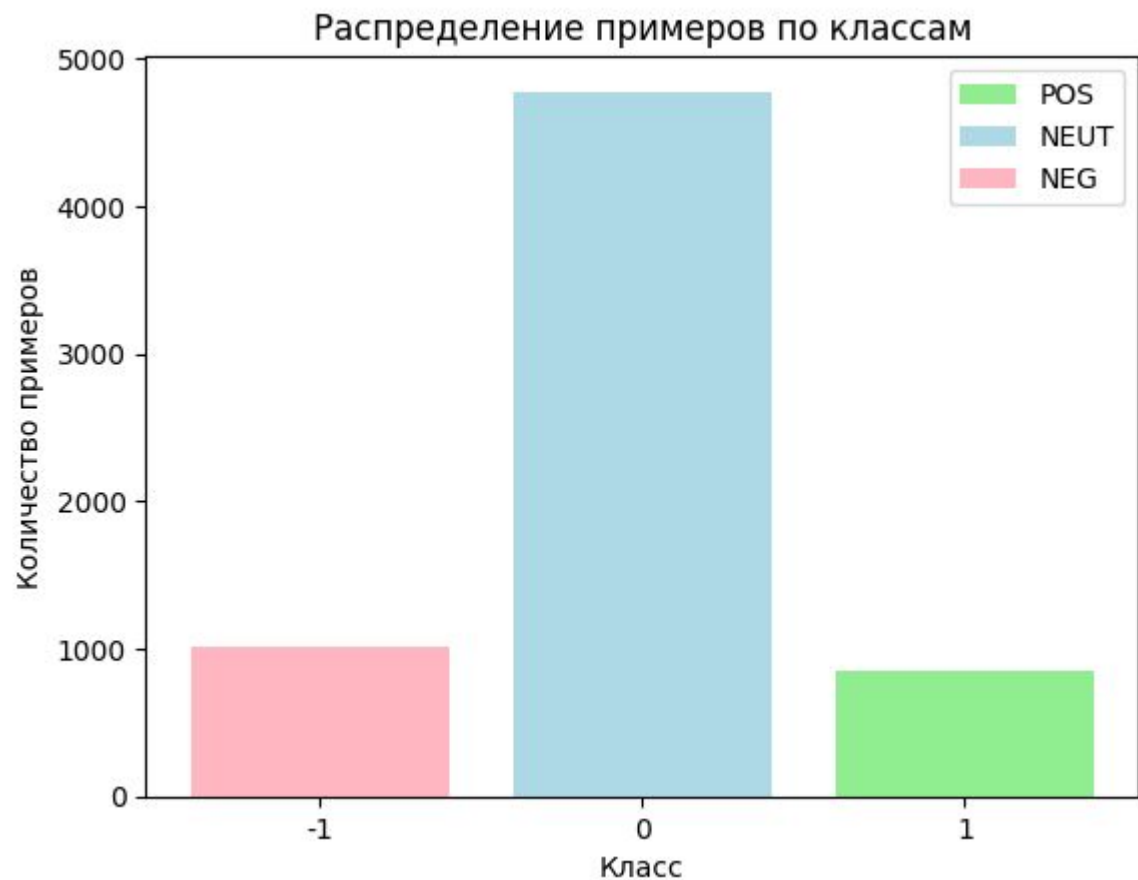
- Обучающая (6637), валидационная (2845) и тестовая (1947) выборки
- -1 – негативная тональность, 0 – нейтральная, 1 – позитивная

	sentence	entity	entity_tag	entity_pos_start_rel	entity_pos_end_rel	label
0	Джеймс «Бадди» Макгирт (James (Buddy) McGirt, ...	спортсмена	PROFESSION	86	96	0
1	«За всю нашу долгую карьеру нам довелось играт...	музыкантов	PROFESSION	258	268	0
2	Ранее, 7 декабря, толпа болельщиков перекрыла ...	Егоре Свиридове	PERSON	175	190	0
3	В субботу, 21 июля 2018 года, на арене СК «Оли...	россиянина	NATIONALITY	122	132	0
4	Представитель талибов Забиулла Муджахид в твит...	американских захватчика	PROFESSION	109	132	-1

- Длина большинства текстов находится в диапазоне от **6** до **28** слов. Минимальная длина текста – **3**, максимальная – **62**; средняя длина текста – **20,5**.



- Явный недобор сущностей-национальностей и перевес нейтральных примеров



Планируемая модель

- *Гипотеза*: **русскоязычные** LLM с задачей справятся **лучше мультязычных**;
- *Метод*: **zero-shot** промптинг LLM, возможен **few-shot** в случае осмысленности попытки через него улучшить качество;
- Проблема дисбаланса классов будет решена **модификацией функции потерь** с помощью **взвешивания**;
- **Фиксированный** промпт.