

О ЧЕМ КУРС?

Цель курса – это интеграция знаний, полученных в ходе изучения теоретических и компьютерных дисциплин в единую систему, развитие практических навыков работы с современными инструментами компьютерной лингвистики, формирование опыта научно-исследовательской работы.

Задачи курса:

1. Научиться выявлять актуальные проблемы в области компьютерной лингвистики и формулировать проектные задачи.
 2. Пройти полный цикл реализации проектов: от первоначальной предобработки данных до анализа результатов и их интерпретации.
 3. Уделить внимание работе с большими наборами данных и анализу текстовой информации с использованием современных инструментов.
- Для примера были предложены темы проектов из соревнований конференции «Диалог» 2022, 2023 и 2024 года.
 - При желании можно было предложить свою тему.



АНАЛИЗ ТОНАЛЬНОСТИ К ИМЕНОВАННЫМ СУЩНОСТЯМ В НОВОСТНЫХ ТЕКСТАХ

*Курс «Проектные задачи
компьютерной лингвистики»*

*Преподаватель Студеникина
Ксения Андреевна*

СОРЕВНОВАНИЕ ПО АНАЛИЗУ ТОНАЛЬНОСТИ К ИМЕНОВАННЫМ СУЩНОСТЯМ В НОВОСТНЫХ ТЕКСТАХ

RuSentNE-2023 (Golubev et al. 2023)

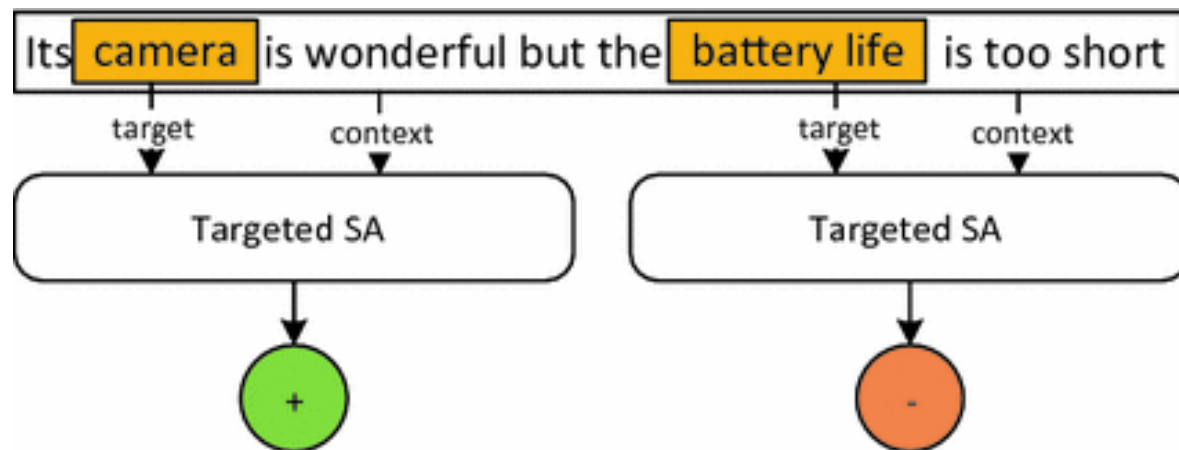
Таргетированный анализ тональности (targeted sentiment analysis, TSA) на материале русскоязычных новостных текстов.

Dialogue Evaluation 2023

- Репозиторий на GitHub: <https://github.com/dialogue-evaluation/RuSentNE-evaluation>
- Страница на CodaLab: <https://codalab.lisn.upsaclay.fr/competitions/9538>

ПОСТАНОВКА ЗАДАЧИ

- **Анализ тональности текста** — одно из наиболее широко обсуждаемых направлений в автоматической обработке текстов.
- Активно развивается так называемый **таргетированный** анализ тональности, при котором требуется определять отношение к конкретным **сущностям** и их **свойствам** или **темам**.



ПОСТАНОВКА ЗАДАЧИ

- Особым направлением является таргетированный анализ тональности **новостных текстов**
 - Обнаружение трендов важно для построения **аналитических систем**, отслеживания **имиджа** в СМИ конкретных **людей** или **компаний**.
- Особенности новостных текстов с точки зрения анализа тональности (Loukachevitch, Rusnachenko 2018):
 - содержат мнения, высказанные **разными субъектами**, включая взгляды авторов, позиции цитируемых источников и отношения упомянутых сущностей друг к другу;
 - могут содержать **несколько именованных сущностей** с разными тональностями, что затрудняет определение тональности к конкретной сущности;
 - содержат множество именованных сущностей с **нейтральной тональностью** → нейтральный класс преобладает;
 - значительная доля тональности носит **имплицитный характер** и может быть выведена из действий субъектов.

ПОСТАНОВКА ЗАДАЧИ

Тональность по отношению к сущности в новостном тексте может происходить по крайней мере **из трёх разных источников**:

- мнения **автора** текста,
- **цитируемого** мнения, сам носитель мнения может быть не упомянут в тексте,
- **имплицитного** мнения, которое следует из упомянутых действий или реакций (X уволил Y). Такая информация часто присутствует даже при внешне нейтральном изложении событий.

Пример таргетированного анализа тональности в новостных текстах:

Walmart подала иск в **суд Арканзаса**, в котором находится штаб-квартира компании, в котором говорится, что **Visa** на протяжении нескольких лет завышала межбанковскую комиссию при расчётах по банковским картам.

- Сущность: **суд Арканзаса**, тип: организация, тональность: нейтральная
- Сущность: **Visa**, тип: организация, тональность: негативная

ПРЕДШЕСТВУЮЩИЕ ИССЛЕДОВАНИЯ

- Соревнования на воркшопах SemEval для **английского** и **арабского**:
 - В 2015 г. задача состояла в том, чтобы определить отношение к конкретным характеристикам, обсуждаемым в **отзывах** пользователей, таким как **еда** или **обслуживание** в ресторанах (Pontiki et al., 2015; Pontiki et al., 2016).
 - В 2016-2017 г. требовалось определить отношение к определенной **теме**, обсуждаемой в посте из **социальной сети** (Nakov et al., 2016; Rosenthal et al., 2017).
- Соревнования для **русского** языка:
 - В 2015 г. была изучена оценка тональности на основе обзоров на **рестораны** и **автомобили** (Loukachevitch et al., 2015).
 - В 2015-2016 г. решалась задача мониторинга репутации **банков** и **телекоммуникационных компаний** в постах из **социальной сети** (Loukachevitch and Rubtsova, 2015).

ПРЕДШЕСТВУЮЩИЕ ИССЛЕДОВАНИЯ

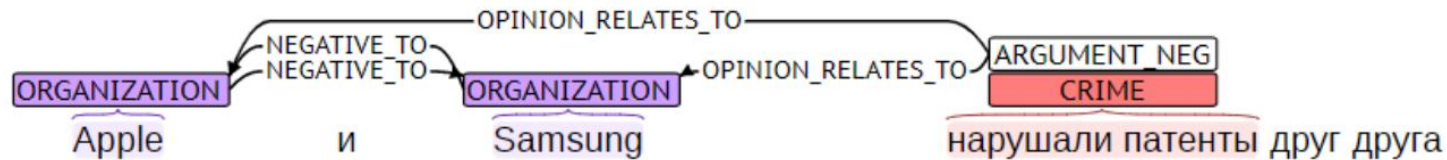
- Большинство международных и российских соревнований посвящены таргетированному анализу тональности применительно к **социальным сетям** и **отзывам** пользователей, а не анализу новостей.
- Общая задача **анализа тональности** на основе цитат из **российских новостей** решалась в ходе соревнования ROMIP-2013 (Chetviorkin and Loukachevitch, 2013).
- Наиболее близкой к интересующей нас задаче является исследование на материале является NewsTSC corpus (Hamborg et al., 2021).
 - Авторы отобрали статьи из 14 **газет**.
 - Были извлечены предложения, в которых присутствуют **именованные сущности**.
 - Разметчики должны определить **отношение** к выделенной сущности.
 - В общей сложности было размечено 3002 предложения, наиболее **частым** было **нейтральное** отношение (2087).

КОРПУС RUSENTNE

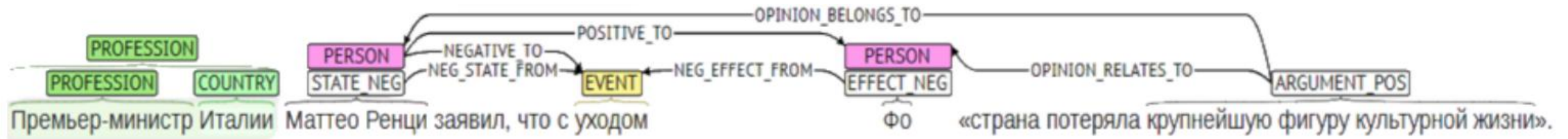
- Исходные данные – датасет NEREL: тексты из Wikinews, размеченные по 29 **именованным сущностям** (Loukachevitch et al., 2021).
- Для отбора текстов использовался **оценочный словарь** RuSentiLex (Loukachevitch and Levchik, 2016).
- Размеченные именованные сущности использовались в качестве **таргетов** для оценки тональности в корпусе RuSentNE.
- Разметка тональности содержит 12 тегов **сущностей** и 11 типов **отношений**, которые описывают тональность, аргументы, эффекты и эмоции сущностей, упомянутых в тексте.

КОРПУС RUSENTNE

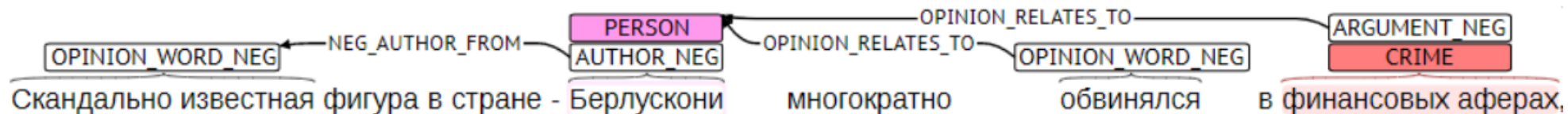
- Две компании – Apple и Samsung – негативно относятся друг к другу (отношение NEGATIVE_TO). Также отмечен отрицательный аргумент, объясняющий отношение двух компаний друг к другу (отношение OPINION_RELATES_TO).



- Маттео Ренци позитивно относится к Фо (отношение POSITIVE_TO). Пояснение отмечено тегом ARGUMENT_POS. Эмоциональное состояние Маттео Ренци негативное из-за смерти Фо. Также размечен негативный эффект для Фо из-за его смерти.



- Позиция автора по отношению к Берлускони обозначена тегом AUTHOR_NEG.



СОРЕВНОВАНИЕ RUSENTNE-2023

Для соревнования RuSentNE-2023 использовалось подмножество аннотированных тегов сущностей и отношений из первоначальной аннотации корпуса RuSentNE:

- теги AUTHOR_POS, AUTHOR_NEG описывают отношение автора к сущности
- отношения POSITIVE_TO, NEGATIVE_TO описывают взаимосвязь между двумя сущностями в тексте
- отношение OPINION_RELATES_TO описывает связь между мнением в тексте и сущностью
 - OPINION_WORD_NEG, ARGUMENT_NEG – негативное
 - OPINION_WORD_POS, ARGUMENT_POS – позитивное

Сущность	Тональность	Источник в RuSentNE
Samsung	негативная	NEGATIVE_TO, ARGUMENT_NEG
Apple	негативная	NEGATIVE_TO, ARGUMENT_NEG
Фо	позитивная	POSITIVE_TO, ARGUMENT_POS
Маттео Ренци	нейтральная	отсутствие разметки
Берлускони	негативная	AUTHOR_NEG, ARGUMENT_NEG

СОРЕВНОВАНИЕ RUSENTNE-2023

Участникам предлагается задача извлечения из новостных текстов

- **тональности** трёх классов: негативная, позитивная, нейтральная
- по отношению к заранее размеченным **сущностям**:
 - PERSON — физическое лицо,
 - ORGANIZATION — организованная группа людей или компания,
 - COUNTRY — нация или территория с единым правительством,
 - PROFESSION — работа, должности в различных организациях и профессиональные звания,
 - NATIONALITY — существительные, обозначающие граждан страны, и прилагательные, соответствующие национальностям.

в рамках отдельного предложения.

ФОРМАТ ДАННЫХ

- sentence – предложение
- entity – объект анализа тональности
- entity_tag – тип сущности: PERSON, ORGANIZATION, PROFESSION, COUNTRY, NATIONALITY
- entity_pos_start_rel – индекс начального символа сущности
- entity_pos_end_rel – индекс следующего символа после последнего символа сущности
- label – метка тональности:
 - 0 – нейтральная,
 - -1 – отрицательная,
 - 1 – положительная

МЕТРИКИ ДЛЯ ОЦЕНКИ

- Основным показателем является макро F_1 -мера (макро F_{1pn} -мера), которая усредняется по двум классам тональности – позитивному (p) и негативному (n).
- Нейтральный класс исключается, поскольку интереснее и важнее извлекать положительно или отрицательно окрашенную тональность.
- Используется следующая процедура:
 - для каждого класса тональности F_1 -мера рассчитывается отдельно;
 - значения F_1 -меры усредняются по двум из трех классов.
- Макро F_{1pn0} -мера для классификации по трем классам является вспомогательной метрикой.

РЕЗУЛЬТАТЫ СОРЕВНОВАНИЯ

Участник	Макро F_{1pn} -мера	Макро F_{1pn0} -мера
mtsai	66.67	74.11
cookies	66.64	74.29
lsanochkin	62.92	71.20
ChatGPT	60.06	70.79
antongolubev	59.64	69.04
sag_m	59.33	68.71
mitrokosta	58.68	67.54
Dmitry315	53.60	62.92
ild	53.20	63.78
abc111	49.98	61.32
Naumov_al	46.96	54.92
baseline_model	40.92	56.71

Подробное описание решений представлено в статье организаторов (Golubev et al. 2023) и других статьях сборника 2023 г.

БАЗОВАЯ МОДЕЛЬ

- Решение организаторов baseline_model (Golubev et al. 2023)
- Модель conversational ruBERT_{base}^{*}
- При тонкой настройке модели задача таргетированного анализа тональности ставилась как классификация пары предложений (Sun et al., 2019)
- На вход подаются два предложения, разделенные токеном [SEP]:
 - Вопрос «Как относятся к X?» где X – сущность в дательном падеже;
 - Текст предложения.
- На выходе ожидается метка класса: -1, 0 или 1.

^{*}<https://huggingface.co/DeepPavlov/rubert-base-cased-conversational>

1 МЕСТО В СОРЕВНОВАНИИ

- Решение mtsai (Podberezko et al. 2023)
- Модели RuRoBERTa_{large}^{*}, XLM-RoBERTa_{large}^{**}, RemBERT^{***}
- Использовались две отдельные модели:
 - одна применялась к исходному предложению,
 - в другую модель подавалось предложение, где сущности заменены токеном [MASK]
- Это помогает модели сосредоточиться на контексте, окружающем сущность.
- В качестве функции потерь использовалась взвешенная кросс-энтропия:
 - примерам с позитивной и негативной тональностью присваивается вес 1,
 - нейтральные предложения имеют вес 0,1.
- Меньший вес нейтральных примеров обусловлен целевой метрикой и преобладанием нейтрального класса в тренировочных данных.

*<https://huggingface.co/ai-forever/ruRoberta-large>

**<https://huggingface.co/FacebookAI/xlm-roberta-large>

***<https://huggingface.co/google/rembert>

2 МЕСТО В СОРЕВНОВАНИИ

- Решение cookies (Glazkova 2023)
- Модель RuRoBERTa_{large}^{*}
- Опробованы разные способы маркирования сущностей:
 - Выделение с помощью пунктуации: Apple и *Samsung* нарушали патенты друг друга.
 - Выделение с помощью токена [NE]: Apple и <|NE|>Samsung <|NE|> нарушали патенты друг друга.
- Для борьбы с дисбалансом применен метод взвешивания классов WINS (Weighted Inverse of Number of Samples):
 - вес считается как величина, обратная частоте класса и нормализованная по всем классам
 - $w_j = \frac{n}{c \cdot n_j}$, n – количество объектов в датасете, c – количество классов, n_j – количество объектов j -го класса

^{*}<https://huggingface.co/ai-forever/ruRoberta-large>

ТЕСТИРОВАНИЕ LLM

- Решение организаторов ChatGPT (Golubev et al. 2023)
- Предложения из набора данных переведены с помощью библиотеки googletrans*
- Использованный промт:
 - What is the attitude of the sentence [translated sentence] to the target [translated entity]?
Select one from «favor, against or neutral» and explain why.
- В нескольких примерах (< 0,1% данных) сущность была переведена неправильно → отсутствовала в предложении. Таким предложениям была присвоена нейтральная метка.

*<https://pypi.org/project/googletrans/>

ТЕСТИРОВАНИЕ LLM

- Продолжение исследования LLM (Rusnachenko et al. 2024):
 - на основе декодера – GPT-4, GPT-3.5, Mistral*, DeciLM**, Microsoft-Phi-2***, Gemma****,
 - на основе энкодера и декодера – Flan-T5*****
- What is the attitude of the author or another subject in the sentence X to the target t? Choose from: positive, negative, neutral.
- Каково отношение автора или другого субъекта в предложении X к t? Выбери из трех вариантов: позитивная, негативная, нейтральная.
- Применен метод промтинга Three-hop Reasoning (THoR) – рассуждение на основе 3-х шагов, который шаг за шагом выявляет имплицитный аспект, мнение и тональность.

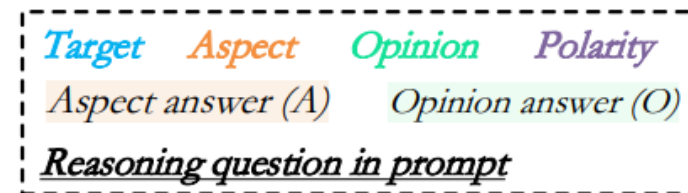
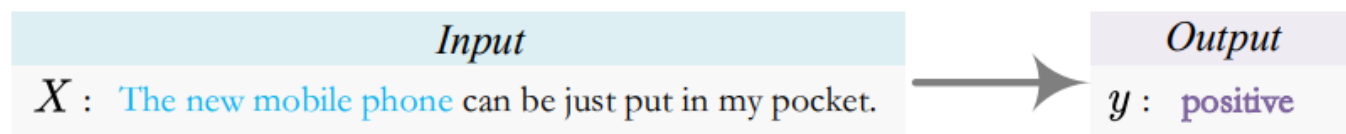
* <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1>, <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

**<https://huggingface.co/Deci/DeciLM-7B-instruct>



***<https://huggingface.co/microsoft/phi-2>

****<https://huggingface.co/google/gemma-7b-it>, <https://huggingface.co/google/gemma-2b-it>







*****<https://huggingface.co/google/flan-t5-xl>, <https://huggingface.co/google/flan-t5-large>, <https://huggingface.co/google/flan-t5-base>



➤ Traditional Prompting

-  Given the sentence "The new mobile phone can be just put in my pocket.", what is the sentiment polarity towards **the new mobile phone**?
-  The sentiment polarity of the given sentence is **neutral**. The sentence simply states that the new mobile phone can be put in the speaker's pocket and does not express any positive or negative sentiment towards the phone.

➤ Three-hop Reasoning with CoT Prompting

- 1st-hop: Inferring aspect**
-  Given the sentence "The new mobile phone can be just put in my pocket.", which specific aspect of **the new mobile phone** is possibly mentioned?
-  The specific aspect of the new mobile phone mentioned in the sentence is **the size or portability**.
- Appending to the next context*
- 2nd-hop: Inferring opinion**
-  Given the sentence, "The new mobile phone can be just put in my pocket.", the specific aspect of the new mobile phone mentioned in the sentence is **the size or portability**. Based on the common sense, what is the implicit opinion towards the mentioned aspect of **the new mobile phone**, and why?
-  Based on the mentioned aspect of size and portability, the sentence implies that the phone is small enough to fit in the speaker's pocket. According to common sense, the implicit opinion of speaker towards the portability is **good**, because the speaker is able to easily carry the phone with them by placing it in their pocket, and find the phone to be **convenient and easy to use**.
- Appending to the next context*
- 3rd-hop: Inferring polarity**
-  Given the sentence, "The new mobile phone can be just put in my pocket.", the specific aspect of the new mobile phone mentioned in the sentence is **the size or portability**. Based on the mentioned aspect of size and portability, the sentence implies that the phone is small enough to fit in the speaker's pocket. According to common sense, the implicit opinion of speaker towards the portability is good, because the speaker can easily carry the phone with them by placing it in their pocket, and find the phone to be convenient and easy to use. Based on such opinion, what is the sentiment polarity towards **the new mobile phone**?
-  The sentiment polarity towards the new mobile phone based on the given sentence is **positive**. The speaker finds the phone to be convenient and easy to use, implying having a favorable view of the phone.

ПРОЕКТ В РАМКАХ КУРСА «НЕЙРОННЫЕ СЕТИ И ИХ ПРИМЕНЕНИЕ В НАУЧНЫХ ИССЛЕДОВАНИЯХ»

- Проект Ю. Соломатиной в рамках курса «Нейронные сети и их применение в научных исследованиях» (Solomatina, Loukachevitch 2024), модель RuRoBERTa_{large}^{*}
- Борьба с дисбалансом классов с помощью аугментации данных позитивного и негативного класса:
 - обратный перевод (на английский язык и обратно),
 - замена 30% токенов на наиболее близкие по косинусному расстоянию
- Метод prompt-tuning: обучение не только весов модели, но и самого промта, объясняющей модели задачу, с помощью библиотеки OpenPrompt^{**}
- {"soft": "Какое отношение выражено к"} {"placeholder": "text_a"} {"soft": "в предложении "} {"placeholder": "text_b"}: {"text": "отрицательное, положительное или нейтрально"}? {"soft": "Отношение в данном предложении к"} {"placeholder": "entity "} {"mask"}
- soft – обучаемые токены, которые инициализируются эмбедингами указанных слов;
text – фиксированные токены, entity– сущность, sentence – предложение

^{*}<https://huggingface.co/ai-forever/ruRoberta-large>

^{**}<https://github.com/thunlp/OpenPrompt>

ЛИТЕРАТУРА-1

- Ilia Chetviorkin and Natalia Loukachevitch. 2013. Evaluating sentiment analysis systems in russian. // Proceedings of the 4th biennial international workshop on Balto-Slavic natural language processing, P 12–17
- Anna Glazkova. 2023. Fine-tuning Text Classification Models for Named Entity Oriented Sentiment Analysis of Russian Texts. // Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”
- Anton Golubev, Nicolay Rusnachenko, and Natalia Loukachevitch. 2023. RuSentNE-2023: Evaluating entityoriented sentiment analysis on Russian news texts. // Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”
- Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning Implicit Sentiment with Chain-of-Thought Prompting. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.
- Felix Hamborg, Karsten Donnay, and Bela Gipp. 2021. Towards target-dependent sentiment classification in news articles. // Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part II 16, P 156–166. Springer.
- Natalia Loukachevitch and Yuliya Rubtsova. 2015. Entity-oriented sentiment analysis of tweets: results and problems. // Text, Speech, and Dialogue: 18th International Conference, TSD 2015, Pilsen, Czech Republic, September 14-17, 2015, Proceedings 18, P 551–559. Springer.
- Natalia Loukachevitch, Pavel Blinov, Evgeny Kotelnikov, Yulia Rubtsova, Vladimir Ivanov, and Elena Tutubalina. 2015. Sentirueval: testing object-oriented sentiment analysis systems in russian. // Proceedings of International Conference Dialog, volume 2, P 3–13.
- Natalia Loukachevitch and Anatolii Levchik. 2016. Creating a general russian sentiment lexicon. // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), P 1171–1176

ЛИТЕРАТУРА-2

- Natalia Loukachevitch and Nicolay Rusnachenko. 2018. Extracting sentiment attitudes from analytical texts. // *Komp'yuternaja Lingvistika i Intellektual'nye Tehnologii*, P 448–458.
- Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Ilia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, and Elena Tutubalina. 2021. Nerel: A russian dataset with nested named entities, relations and events. // *International Conference on Recent Advances in Natural Language Processing: Deep Learning for Natural Language Processing Methods and Applications, RANLP 2021*, P 876–885. Incoma Ltd.
- Pavel Podberezko, Anton Kabaev, Andrey Kaznacheev, and Sabine Abdullayeva. 2023. HAlf-MASked Model for Named Entity Sentiment analysis. // *Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”*
- Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. // *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, P 486–495.
- Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed ALSmadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. // *ProWorkshop on Semantic Evaluation (SemEval-2016)*, P 19–30. Association for Computational Linguistics.
- Nicolay Rusnachenko, Anton Golubev and Natalia Loukachevitch. 2024. Large Language Models in Targeted Sentiment Analysis for Russian. *Lobachevskii Journal of Mathematics* 45.7: 3148-3158.
- Yuliana Solomatina and Natalia Loukachevitch. 2024. Prompt-Tuning Targeted Sentiment Analysis in Russian. *International Conference on Analysis of Images, Social Networks and Texts*. Cham: Springer Nature Switzerland.