

Annotation of Text Corpora by Sentiment and Presence of Irony within a Project of Citizen Science

I. V. Paramonov¹, A. Y. Poletaev¹

DOI: [10.18255/1818-1015-2023-1-86-100](https://doi.org/10.18255/1818-1015-2023-1-86-100)

¹P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 68T50

Research article

Full text in Russian

Received February 3, 2023

After revision February 24, 2023

Accepted February 27, 2023

The paper is devoted to construction of a sentence corpus annotated by the general sentiment into 4 classes (positive, negative, neutral, and mixed), a corpus of phrasemes annotated by the sentiment into 3 classes (positive, negative, and neutral), and a corpus of sentences annotated by the presence or absence of irony. The annotation was done by volunteers within the project “Prepare texts for algorithms” on the portal “People of science”.

The existing knowledge on the domain regarding each task was the basis to develop guidelines for annotators. A technique of statistical analysis of the annotation result based on the distributions and agreement measures of the annotations performed by various annotators was also developed. For the annotation of sentences by irony and phrasemes by the sentiment the agreement measures were rather high (the full agreement rate of 0.60–0.99), whereas for the annotation of sentences by the general sentiment the agreement was low (the full agreement rate of 0.40), presumably, due to the higher complexity of the task. It was also shown that the results of automatic algorithms of detecting the sentiment of sentences improved by 12–13 % when using a corpus for which all the annotators (from 3 till 5) had the agreement, in comparison with a corpus annotated by only one volunteer.

Keywords: sentiment analysis; text corpus; statistical analysis; agreement measures; citizen science

INFORMATION ABOUT THE AUTHORS

Ilya Vyacheslavovich Paramonov
correspondence author

orcid.org/0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org
PhD, associate professor.

Anatoliy Yurievich Poletaev

orcid.org/0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru
post-graduate student.

Funding: The research was performed within the YarSU citizen science project No. CS-02/2022.

For citation: I. V. Paramonov and A. Y. Poletaev, “Annotation of Text Corpora by Sentiment and Presence of Irony within a Project of Citizen Science”, *Modeling and analysis of information systems*, vol. 30, no. 1, pp. 86–100, 2023.

Разметка корпусов текстов по тональности и наличию иронии в рамках проекта гражданской науки

И. В. Парамонов¹, А. Ю. Полетаев¹

DOI: [10.18255/1818-1015-2023-1-86-100](https://doi.org/10.18255/1818-1015-2023-1-86-100)

¹Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912

Научная статья

Полный текст на русском языке

Получена 3 февраля 2023 г.

После доработки 24 февраля 2023 г.

Принята к публикации 27 февраля 2023 г.

Статья посвящена построению корпуса предложений, размеченных по общей тональности на 4 класса (положительный, отрицательный, нейтральный, смешанный), корпуса фразеологизмов, размеченных по тональности на 3 класса (положительный, отрицательный, нейтральный), и корпуса предложений, размеченных по наличию или отсутствию иронии. Разметку проводили волонтеры в рамках проекта «Готовим тексты алгоритмам» на портале «Люди науки».

На основе имеющихся знаний о предметной области для каждой из задач были составлены инструкции для разметчиков. Также была выработана методика статистической обработки результатов разметки, основанная на анализе распределений и показателей согласия оценок, выставленных разными разметчиками. Для разметки предложений по наличию иронии и фразеологизмов по тональности показатели согласия оказались достаточно высокими (доля полного совпадения 0.60–0.99), при разметке предложений по общей тональности согласие оказалось слабым (доля полного совпадения 0.40), по-видимому, из-за более высокой сложности задачи. Также было показано, что результаты работы автоматических алгоритмов анализа тональности предложений улучшаются на 12–13 % при использовании корпуса, относительно предложений которого сошлись мнения всех разметчиков (3–5 человек), по сравнению с корпусом с разметкой только одним волонтером.

Ключевые слова: анализ тональности; текстовый корпус; статистический анализ; показатели согласия; гражданская наука

ИНФОРМАЦИЯ ОБ АВТОРАХ

Илья Вячеславович Парамонов
автор для корреспонденции

orcid.org/0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org
канд. физ.-мат. наук, доцент.

Анатолий Юрьевич Полетаев

orcid.org/0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru
аспирант.

Финансирование: Исследование выполнено в рамках проекта гражданской науки ЯРГУ № CS-02/2022.

Для цитирования: I. V. Paramonov and A. Y. Poletaev, “Annotation of Text Corpora by Sentiment and Presence of Irony within a Project of Citizen Science”, *Modeling and analysis of information systems*, vol. 30, no. 1, pp. 86–100, 2023.

Введение

Построение размеченных текстовых корпусов является актуальной задачей компьютерной лингвистики, поскольку предоставляет исходные данные практически для любых исследований в данной области. Решение данной задачи является чрезвычайно трудозатратным, так как достижение высокого качества построенных корпусов требует большого количества работы, выполняемой вручную. Поскольку критерии разметки часто оказываются трудноформализуемыми (например, в случае разметки текста по тональности), желательным является выполнение разметки экспертами в предметной области, что ещё больше увеличивает стоимость выполнения данного вида работ.

В связи с высокой трудоёмкостью исследователи часто прибегают к автоматической или полуавтоматической разметке текстов (например, разметке текстовых отзывов на товары по их рейтингу, задаваемому пользователями на маркетплейсах), а также привлечению неквалифицированных разметчиков-волонтеров. В обоих случаях качество корпусов будет ниже, чем в случае ручной разметки экспертами, однако построенные таким образом корпуса всё же могут быть полезны в различных исследованиях после агрегации собранных данных и хотя бы избирательной их верификации [1].

Не существует общей методики составления инструкций по разметке текстовых корпусов для волонтеров, а также обработки результатов разметки. Составление инструкций — достаточно сложная задача, т. к. они, с одной стороны, должны быть достаточно полными, чтобы неквалифицированные разметчики могли в подавляющем большинстве случаев ориентироваться на них, а с другой стороны, не должны быть слишком сложными и объёмными, чтобы избежать сложностей с их пониманием. От качества методики обработки результатов разметки напрямую зависит применимость полученного корпуса для решения задач компьютерной лингвистики [2].

Настоящая работа посвящена построению корпуса предложений, размеченных по общей тональности на 4 класса (положительный, отрицательный, нейтральный, смешанный), корпуса фразеологизмов, размеченных по тональности на 3 класса (положительный, отрицательный, нейтральный), и корпуса предложений, размеченных по наличию или отсутствию иронии. Разметку проводили волонтеры, набранные в рамках проекта «Готовим тексты алгоритмам» на портале «Люди науки»¹.

1. Обзор связанных работ

В современной научной литературе крайне мало исследований, посвящённых разметке текстовых корпусов и агрегации результатов этой разметки. В большинстве работ, где идёт речь о построении размеченных текстовых корпусов, авторы очень поверхностно описывают или не описывают вообще, как именно они собирались. Можно выделить два основных аспекта, связанных с данной задачей: разработка инструкций по разметке и собственно методика разметки и агрегации её результатов.

Нужно отметить, что существует несколько общих методик агрегации экспертных оценок, например, метод Делфи, однако из-за достаточно высокой трудоёмкости они слабо применимы для агрегации оценок большого числа предложений, выставленных большим числом разметчиков.

Одной из наиболее значимых работ в области построения инструкций для разметки текстов по тональности является статья [3]. В ней отмечается потенциальная сложность задачи, вызванная нечётким определением самого понятия тональности, приводится классификация предложений, сложных для разметки, и предлагаются два подхода к выработке инструкций для разметчиков, предположительно позволяющие повысить точность результатов в сложных случаях. В первом случае используется опросник, содержащий утверждения, характеризующие те или иные классы тональности предложений, а также примеры типов предложений, относящихся к заданному классу (восхищение, поддержка, симпатия и т. п.). Во втором случае разметчикам предлагается отвечать

¹<https://citizen-science.ru/projects/gotovim-teksty-algoritmam.html>

на вопросы относительно эмоционального состояния автора предложения, объекта, по отношению к которому выражена тональность, и характера воздействия предложения на большинство людей. Отмечается, что среди этих двух подходов нет наилучшего, и эффективность каждого из них может определяться решаемой задачей и квалификацией разметчиков.

В другой работе того же автора [4] рассматривается задача разметки твитов по отношению автора к конкретным персоналиям или явлениям (за, против, другое) и о том, высказывается ли мнение об объекте напрямую. Каждый твит рассматривался как минимум восемью разметчиками, что является очень большим значением среди всех существующих работ в данной области. Всего было собрано 5412 твитов; 5 % твитов были изначально размечены их авторами, и если разметчик ошибался при разметке такого твита, он получал уведомление об этом, а если он ошибался более чем на 30 % таких твитов в совокупности, он исключался из работы как недобросовестный. При агрегации результатов брались только такие оценки, с которыми были согласны как минимум 60 % работавших с ними разметчиков. Степень согласия составила примерно 81.85 % для задачи определения отношения автора к персоналии или явлению и 68.9 % для задачи определения того, выражено ли мнение об объекте напрямую. Степень согласия определялась как среднее мер согласия по каждому твиту, которая, в свою очередь, вычислялась как отношение количества разметчиков, давших наиболее распространённый ответ для данного твита, к общему количеству разметчиков этого твита.

В статье [5] используется подход из работы [3] с некоторыми изменениями. В частности, рассматриваются только предложения из 5–15 слов. Кроме того, идея проверки следования разметчика инструкциям, решаемая в оригинальной работе с использованием вопросов с известными ответами, о которых разметчик заранее не знал, заменена выдачей участникам перед началом работы размеченного экспертами набора предложений в качестве примера. Каждое предложение размечалось минимум двумя разметчиками. В случае их согласия третий не привлекался, в случае несогласия — привлекался. Если и третий не давал прийти к решению, привлекались ещё двое. Данные некоторых разметчиков были исключены из рассмотрения, поскольку они допускали слишком много ошибок на первом этапе. Всего было размечено 15744 предложения. Для оценки согласия использовалась альфа Криппендорфа [6], показывающая насколько согласие между разметчиками близко к полному. Её значение получилось равным 0.66, что может говорить о существенном согласии. Также приведены 4 примера сложных для анализа предложений, сходных с описанными в работе [3].

Сходная методика используется в статье [7]: разметка производилась двумя экспертами; в случае, когда оценки не сходились, решение принимал третий разметчик. Всего было размечено 17572 предложения на китайском языке. Для оценки согласия использовалась каппа Коэна [8]. Её значение оказалось равным 0.71–0.73, что считается удовлетворительным.

В другой работе также для китайского языка [9] размечался корпус отзывов на рестораны по различным аспектам (качество еды, обслуживания, цена и т. д.). Для каждого аспекта разметчикам предлагалось ответить только на один вопрос — об отношении автора отзыва к аспекту (положительном, отрицательном или нейтральном). Каждое предложение размечалось двумя независимыми разметчиками, прошедшими краткое обучение; далее проводилась валидация результатов их разметки третьим разметчиком (если оценки сошлись) или экспертом (в противном случае). Спорные случаи также обрабатывались экспертом. В итоге было размечено 46730 предложений. Характеристики согласия не вычислялись.

В работе [10] производилась разметка корпуса твитов, относящихся к определённым брендам. Каждый твит размечался тремя разметчиками, которые должны были указать, какие эмоции из заданного набора (доверие, счастье, грусть и т. д.) присутствуют в заданном твите. Для оценки согласия всех разметчиков по категориям использовалась каппа Флейса [11], для оценки попарного согласия между разметчиками — каппа Коэна. Обе эти метрики характеризуют долю совпавших оценок разметчиков, которая не может быть объяснена случайным совпадением мнений. Значения метрик

составили 0.372 и 0.354 соответственно, что говорит об умеренном согласии разметчиков. Анализ причин несогласия разметчиков не проводился.

В работе [12] рассматривается задача разметки по тональности корпуса постов из социальной сети «ВКонтакте» по социально-политической тематике с вариантами ответа: «положительное», «отрицательное», «не тональное», «сомнительное», «речевой акт» (поздравления, технические сообщения и т. д.). Размечались только посты, содержащие от 10 до 800 символов. К разметке были привлечены 6 экспертов в области лингвистики, которым были выданы соответствующие инструкции. В итоге были размечены 31185 постов, каждый пост размечался тремя разметчиками. Для оценки согласия использовалась каппа Флейса, величина которой составила 0.58.

Из рассмотренных работ видно, что в каждом конкретном случае разметка корпусов осуществлялась по-разному и использовались различные метрики для оценки согласия. Есть и общие черты: практически во всех работах каждый текстовый фрагмент размечается минимум тремя разметчиками (в некоторых случаях ограничиваются даже двумя оценками, если разметчики согласны между собой); используются некоторые инструкции для разметчиков, весьма схожие в разных работах; предпринимаются действия по очистке размеченного корпуса от сомнительных результатов разметки вплоть до удаления результатов тех разметчиков, которые представляются недобросовестными. В большинстве перечисленных работ основным показателем качества разметки считаются метрики согласия. Возможно, это связано с тем, что другие методы анализа, например, верификация экспертами, слишком трудоёмки. Однако степень согласия между разметчиками оказывается достаточно низкой, что, по-видимому, обусловлено нечёткостью понятия тональности и субъективным его восприятием разными людьми.

2. Методика разметки корпусов

В качестве исходного корпуса для разметки предложений по тональности и по наличию иронии был использован OpenCorpora² — открытый корпус предложений из новостных и публицистических текстов на русском языке, относящихся к различным предметным областям. Для задачи разметки предложений по тональности использовались только предложения из семи и более слов. Список фразеологизмов и их значений для разметки по тональности был взят из Викисловаря³.

Перед разметчиками ставилась задача оценить с точки зрения тональности или наличия иронии предложения или фразеологизмы из индивидуального набора. Наборы формировались автоматически таким образом, чтобы, во-первых, каждый разметчик оценивал каждое предложение не более одного раза, а во-вторых, каждое предложение или фразеологизм оказались бы в комплектах хотя бы трёх разметчиков. В комплектах для разметки фразеологизмов для удобства разметчиков были также приведены словарные значения фразеологизмов.

Для разметчиков был проведён инструктаж, во время которого их ознакомили с предметной областью и руководствами по разметке.

В руководстве для задачи разметки предложений по тональности было указано, что тональность предложения показывает, как автор относится к теме предложения. Были выделены четыре класса тональности предложений и класс предложений с неопределимой (сомнительной) тональностью и приведены их отличительные черты, а также примеры предложений каждого из классов. Далее приведена выдержка из руководства в части описания различных классов.

- Положительные предложения, в которых автор выносит положительные оценки или выражает положительные эмоции. В положительном предложении могут упоминаться отрицательные оценки или эмоции, но автор должен явно указать, что их не разделяет. Примеры положительных предложений:

²<https://opencorpora.org>

³<https://ru.wiktionary.org>

Листьев достиг той раскованности в кадре, о какой раньше никто не помышлял.

На его взгляд, руководство, организовавшее компанию по достижению цели, заслуживает самой высокой похвалы.

Мне очень понравился фильм, а тот, кто говорит, что он скучный, просто ничего не понял.

- Отрицательные предложения, в которых автор выносит отрицательные оценки или выражает отрицательные эмоции. Положительно характеризующая тему информация отсутствует, либо является несущественной. Примеры отрицательных предложений:

В целом получилось так, что проиграли все.

Выяснилось, что отверстие водостока в шикарном душевой кабинке забито и поэтому принять душ было совершенно невозможно.

- Предложения со смешанной тональностью, в которых автором выносятся как положительные, так и отрицательные оценки и выражаются эмоции, соответствующие авторской позиции. Данный класс был выделен специально для предложений, являющихся тональными (выражающих авторские эмоции и содержащих авторские оценки), при этом не относимых однозначно к положительным или отрицательным. Примеры таких предложений:

Героев «Чайки», слабых и не слишком замечательных, но всё-таки порядочных людей, Акунин изобразил порочными и преступными.

Сделка между Google и RM получила летом хорошее освещение в медиа, но не стала сенсацией.

Виктор Пелевин пишет не в пример хуже Акунина, но у Пелевина есть что сказать читателю.

- Нейтральные предложения, в которых автор не выносит своих оценок и не выражает согласия с чужими. Нейтральные предложения безэмоциональны, в них просто «сухо» сообщается о некоторых фактах, авторское же отношение к теме не проявляется. Примеры нейтральных предложений:

Информация о том, что фильм провалился в прокате, не подтвердилась.

По традиции страна-хозяйка мирового первенства освобождается от отборочного турнира.

Газета «Пражский экспресс» утверждает, что актёры вели себя корректно, спектакль срывать не собирались, а лишь высказали требование о выплате гонораров.

- Предложения с неопределимой (сомнительной) тональностью, мысль автора в которых неясна, например, из-за того, что предложение неполное и без контекста не имеет смысла. Примеры таких предложений:

Автор вопроса уверен. . . , или история повторяется.

Ведь русские могли отходить до Урала.

Потому что это чтение с самого начала слишком музыка.

Также руководство содержало ряд общих рекомендаций:

- считать положительное либо отрицательное отношение автора вне контекста и возникающих ассоциаций;
- относить к классу смешанной тональности предложения, в которых авторское отношение однозначно есть, но неясно, положительное оно или отрицательное;

- относить к нейтральным целостные предложения, содержащие законченную мысль, но не выражающие авторское отношение; если же мысль не закончена или предложение не целостно, то считать тональность предложения неопределимой;
- при сомнениях ориентироваться на эмоциональную составляющую предложений;
- при возникновении вопросов не обсуждать их с другими разметчиками, чтобы не влиять на их восприятие.

Руководство для задачи разметки фразеологизмов по тональности было составлено аналогично; основное отличие состояло в том, что были оставлены только три класса тональности — фразеологизмы с положительной тональностью, фразеологизмы с отрицательной тональностью и нейтральные фразеологизмы. Класс смешанной тональности не использовался из-за того, что фразеологизмы гораздо короче предложений, и в них не может проявляться одновременно как положительное, так и отрицательное отношение. По схожей причине не использовался и класс неопределимой тональности — поскольку фразеологизмы достаточно коротки, а авторское отношение в них проявляется достаточно явно, все случаи, когда непонятно, положительное отношение проявляет автор или отрицательное, относятся к нейтральному классу.

Руководство для задачи разметки предложений по наличию иронии содержало примеры предложений, содержащих иронию. Также перед началом выполнения задачи разметчикам требовалось изучить определение иронии из словаря лингвистических терминов под редакцией Т. В. Жеребило [13].

Характеристики полученных данных приведены в таблице 1. Значительный размах числа оценок каждого разметчика в задачах разметки фразеологизмов по тональности и предложений по иронии обусловлен тем, что некоторые разметчики не выполнили предложенные им задания до конца.

Table 1. Corpora and annotation characteristics

Таблица 1. Характеристики корпусов и их разметки

Тип разметки	Объём корпуса	Число разметчиков	Число классов	Число оценок каждого элемента	Число оценок каждого разметчика
предлож./тональность	11608	18	5	3–5	2214–2218
фразеолог./тональность	1253	12	3	3–5	140–354
предлож./ирония	35013	14	2	1–3	1922–4000

3. Методика статистической обработки данных

Для исследования выставленных разметчиками оценок использовалась следующая методика.

1. Изучалось распределение оценок по классам у разных разметчиков. Поскольку предложения и фразеологизмы распределялись случайно, а объём индивидуальных комплектов для разметки был достаточно велик, можно ожидать, что при одинаковом понимании всеми разметчиками задачи доли оценок каждого класса для всех разметчиков будут приблизительно одинаковы. Следовательно, если доли оценок какого-либо класса у разметчиков значительно отличаются, это даёт основания считать, что разметчики решали задачу по-разному. Исследование распределения оценок по классам проводилось графическим методом на основе построенных для каждого класса гистограмм распределения доли оценок этого класса и диаграмм размаха. Также с помощью критерия хи-квадрат было проверено, совпадает ли распределение оценок по классам у каждого из разметчиков с распределением по классам всех оценок всех разметчиков. Несовпадение распределений также даёт основания считать, что способ выполнения задания у разметчика значительно отличается от способов остальных разметчиков.

2. Оценивалось согласие оценок, выставленных разными разметчиками. Если мнения разных разметчиков относительно класса одного предложения или фразеологизма часто расходятся, это может свидетельствовать как о различном понимании разметчиками задачи, так и о затруднениях в единообразном следовании инструкциям, например, из-за слишком высокой роли индивидуальных особенностей восприятия в процессе определения тональности. Согласие оценивалось на основе двух показателей: доли предложений, для которых совпали мнения всех оценивавших их разметчиков, и альфы Криппендорфа. Альфа Криппендорфа показывает, насколько согласие между разметчиками близко к полному, и рассчитывается по формуле

$$\alpha = \frac{A_o - A_e}{1 - A_e},$$

где A_o — количество совпавших оценок разметчиков; A_e — количество оценок, которые совпали бы, если бы разметчики выставляли оценки случайно [14]. В работе [15], выделяют следующие эмпирические границы для различных степеней согласия:

- $\alpha \leq 0.2$ — согласие практически отсутствует;
- $0.2 < \alpha \leq 0.4$ — слабое согласие;
- $0.4 < \alpha \leq 0.6$ — умеренное согласие;
- $0.6 < \alpha \leq 0.8$ — существенное согласие;
- $\alpha > 0.8$ — практически полное согласие.

Нужно отметить, что альфа Криппендорфа учитывает различия в оценках нескольких разметчиков, а не только их полное совпадение. Например, она будет выше в случае, когда среди трёх разметчиков оценки только двух совпали, чем в случае, когда все три разметчика дали разные оценки.

3. Для разметки предложений по тональности оценивалось, насколько сильно улучшится согласие, если отбросить предложения, тональность которых была отмечена как неопределимая хотя бы одним разметчиком, и предложения, тональность которых хотя бы один разметчик оценил как положительную, а другой — как отрицательную. Значительное (более, чем на 0.05) улучшение показателей согласия даст основания считать, что существенная часть несовпадающих оценок разметчиков относится именно к таким предложениям. Поскольку в речи не должно быть слишком много как предложений, тональность которых определить невозможно, так и предложений, которые одному человеку кажутся положительными, а другому отрицательными, если в разметке таких случаев оказывается много, то, возможно, руководства по разметке содержат некорректное или плохо сформулированное понятие тональности.

4. Результаты статистической обработки

4.1. Результаты обработки разметки предложений по тональности

По гистограммам распределения долей оценок каждого класса оценок и диаграмме размаха (рис. 1) явных выбросов, которые бы свидетельствовали о том, что один из разметчиков решал задачу определения предложений некоторого класса существенно иначе, чем остальные, не наблюдается. У всех разметчиков доли оценок положительного и отрицательного классов и класса смешанной тональности достаточно близки, с медианами 0.15 для положительной, 0.30 для отрицательной и 0.10 для смешанной тональности. Доля нейтральных оценок варьируется достаточно сильно — от 0.30 до 0.80. Доля предложений, тональность которых определить не получилось, изменяется ещё сильнее: у половины разметчиков она не превышает 0.01, однако у другой половины практически равномерно распределена на промежутке от 0.01 до 0.07, т. е. некоторые разметчики значительно чаще считали тональность предложения неопределимой, чем остальные. По-видимому, эти разметчики чувствовали неуверенность при решении задачи.

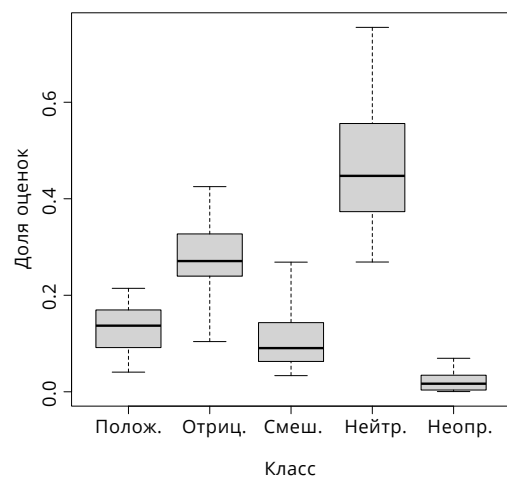
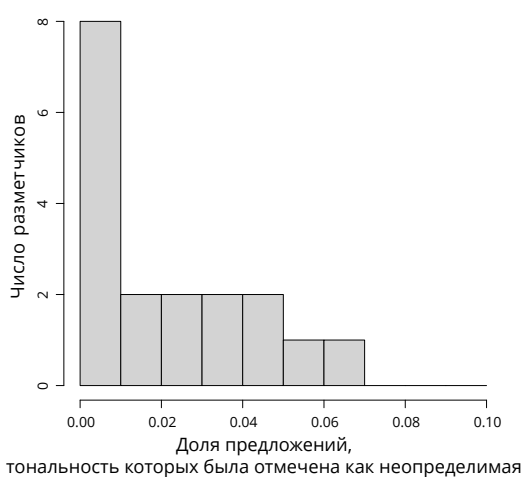
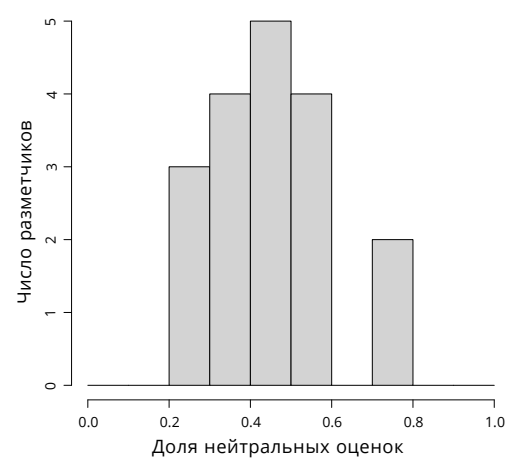
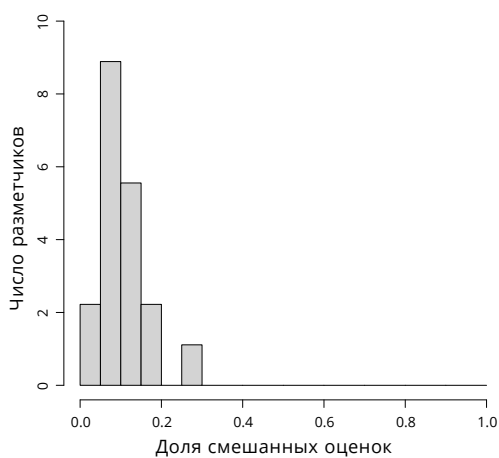
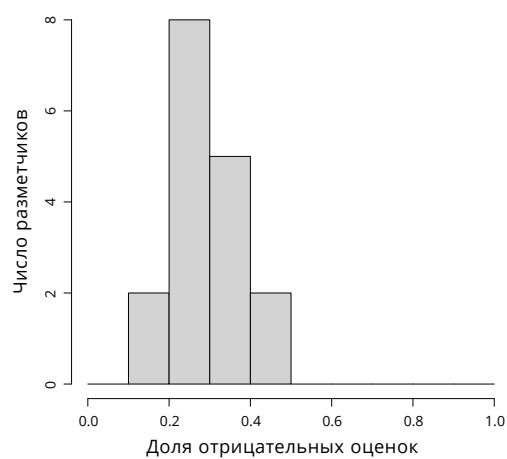
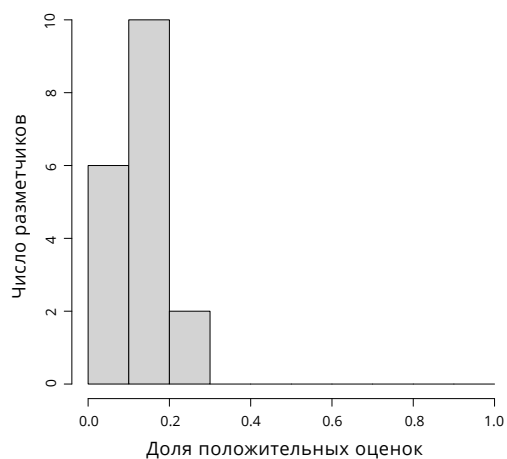


Fig. 1. Distribution of marks classes in the sentences sentiment markup

Рис. 1. Распределение классов оценок в разметке предложений по тональности

Для каждого из разметчиков критерий хи-квадрат показывает наличие значимых различий между распределениями по классам оценок этого разметчика и оценок всех остальных разметчиков. Следовательно, есть основания считать, что способы выполнения задания у разных разметчиков существенно отличались друг от друга, в первую очередь — из-за разного понимания нейтральной тональности.

Доля предложений, для которых совпали мнения всех оценивавших их разметчиков, составила 0.40, что является признаком достаточно слабого согласия в оценках разметчиков. Альфа Кrippендорфа составила 0.35, что также соответствует слабому согласию.

Предложений, тональность которых хотя бы один разметчик отметил как неопределимую, оказалось 831 (7 % всех предложений). После их удаления доля предложений, для которых совпали мнения всех оценивавших их разметчиков, увеличилась на 0.03, до 0.43; альфа Кrippендорфа также увеличилась на 0.03, до 0.38. Кроме того, в корпусе оказалось 338 предложений, отмеченных хотя бы одним разметчиком как имеющие положительную тональность, и хотя бы одним — как отрицательную (3 % всех предложений). После их удаления осталось 10454 предложения, т. к. тональность некоторых предложений разными разметчиками была отмечена и как неопределимая, и как положительная, и как отрицательная. Доля предложений, для которых совпали мнения всех оценивавших их разметчиков, увеличилась на 0.01, до 0.44, а альфа Кrippендорфа — на 0.02, до 0.40. Общее увеличение доли предложений, для которых совпали мнения всех разметчиков, составило 0.04, а альфы Кrippендорфа — 0.05. В итоге показатели согласия повысились на 0.04–0.05, и согласие осталось достаточно слабым.

4.2. Результаты обработки разметки фразеологизмов по тональности

На гистограммах распределения долей оценок каждого класса и диаграмме размаха (рис. 2) явных выбросов не наблюдается. Доля положительных оценок для всех разметчиков приблизительно одинакова, в среднем составляет 0.20. Доли отрицательных и нейтральных оценок изменяются сильнее — от 0.30 до 0.70 и от 0.10 до 0.60 соответственно. Особенно стоит отметить распределение доли нейтральных оценок: при медиане 0.35 трое из двенадцати разметчиков посчитали нейтральными менее 0.2 оцениваемых фразеологизмов. Такая аномалия свидетельствует о том, что некоторые разметчики, по-видимому, определяли фразеологизмы нейтральной тональности существенно иначе, чем остальные.

Проверка совпадения распределений оценок по классам с помощью критерия хи-квадрат показала, что для 4 из 12 разметчиков нет значимых различий между распределением оценок данного разметчика и оценок всех остальных разметчиков, а для 8 из 12 разметчиков есть.

Доля фразеологизмов, для которых совпали мнения всех оценивавших их разметчиков, составила 0.60; альфа Кrippендорфа — 0.54, что соответствует умеренному согласию.

Следовательно, задача разметки фразеологизмов по тональности была понята и выполнялась разметчиками гораздо более единообразно, чем задача разметки предложений по тональности. Об этом свидетельствуют и более единообразные распределения оценок разметчиков по классам, и значительно более высокие показатели согласия. В то же время согласие между разметчиками далеко от полного. Можно было ожидать, что тональность фразеологизмов будет выражаться достаточно ярко и более определённо по сравнению с полными предложениями, однако относительно большого их числа разметчики всё же не сошлись во мнениях.

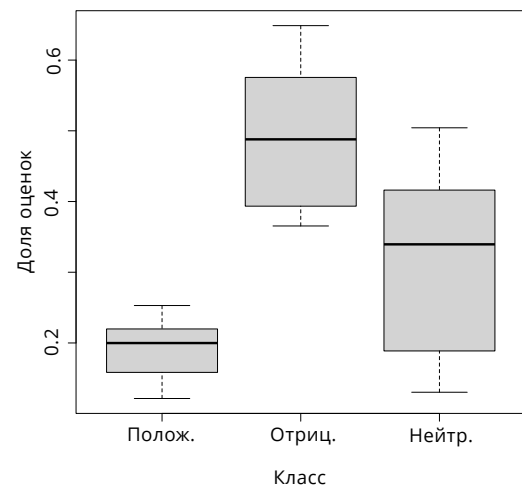
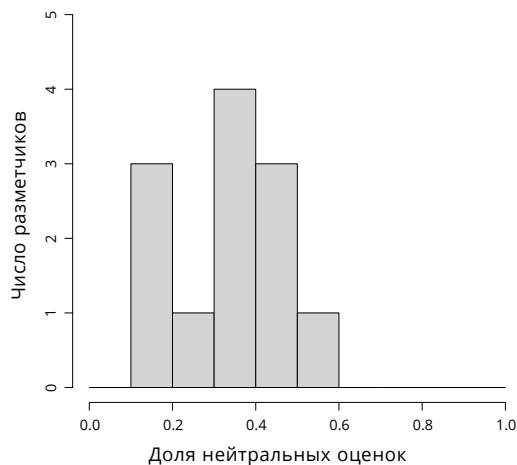
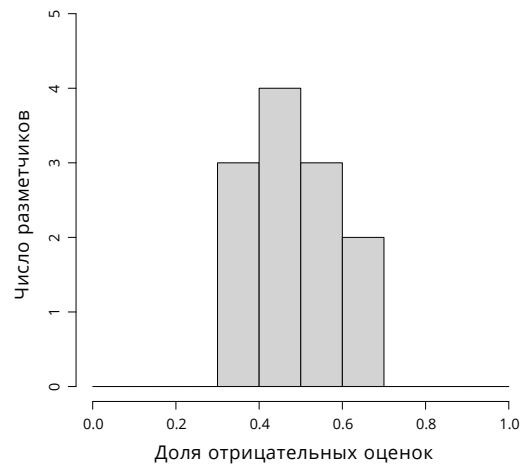
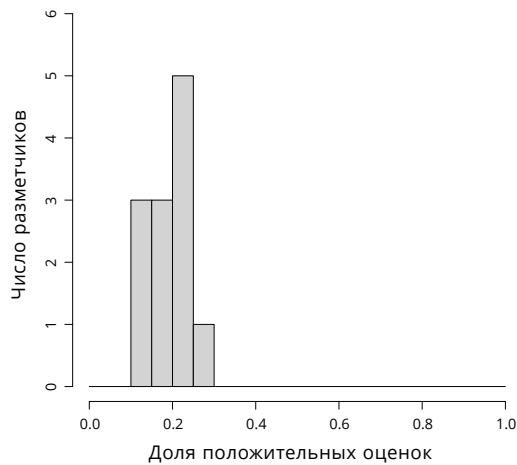


Fig. 2. Distribution of marks classes in the phrasemes sentiment markup

Рис. 2. Распределение классов оценок в разметке фразеологизмов по тональности

4.3. Результаты разметки предложений по наличию иронии

На гистограмме распределения доли предложений, отмеченных как содержащие иронию (рис. 3), явно видны три разметчика, которые отмечали иронию в экстремально высоком числе предложений (0.26, 0.33 и 0.50). Экспертная валидация подтвердила, что их разметка не соответствует инструкциям.

Чтобы проверить, насколько значительной проблемой может быть присутствие в корпусе такой некорректной разметки, были рассчитаны показатели согласия для оценок всех разметчиков, а также всех разметчиков, за исключением тех, кто не следовал инструкциям. Во втором случае доля предложений с полным совпадением оценок увеличилась на 0.06 и достигла 0.99. Учитывая, что альфа Кrippендорфа в данном случае не показательна из-за слишком сильного дисбаланса классов [16], можно говорить о том, что удаление оценок разметчиков, не следовавших предложенным инструкциям, значительно улучшило согласие.

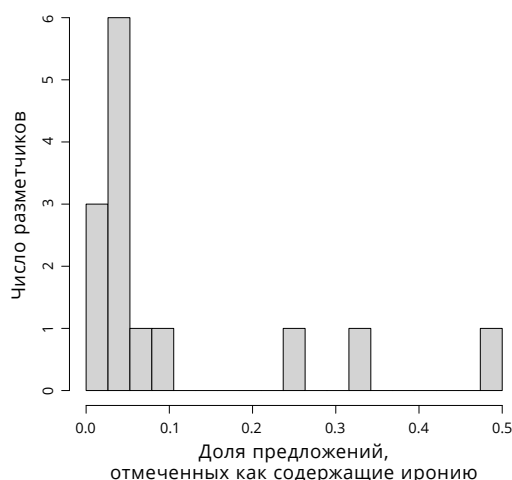


Fig. 3. Distribution of share of irony marks

Рис. 3. Распределение доли предложений, отмеченных как содержащие иронию

5. Обсуждение результатов

Обработка результатов разметки показала, что согласие между разметчиками сильно зависит от сложности задачи. Для наиболее сложной задачи разметки предложений по тональности согласие оказалось достаточно слабым, а для более простой задачи разметки фразеологизмов по тональности — значительно более сильным. Наилучшим же оказалось согласие в задаче разметки предложений по наличию иронии, которая была самой простой из выполненных разметчиками.

Одной из причин несогласия между разметчиками может быть несовершенство инструкций по разметке. Например, тональность предложения «А ежели кому кажется, что я ханжа и зануда, то так оно и есть» по формальным признакам, приведённым в инструкции, может быть определена как отрицательная, так как есть согласие автора с отрицательной оценкой и явно считываются отрицательные эмоции, хотя правильнее будет отметить её как неопределимую, поскольку в предложении присутствует ирония. Встретив такое предложение в тексте, часть разметчиков будет ориентироваться на формальные инструкции, а часть — на здравый смысл, что в итоге приведёт к несогласию. Следовательно, можно попробовать улучшить согласие, точнее сформулировав решаемую задачу и улучшив инструкции. Результаты анализа разметки показывают, что для более простой задачи определения иронии уже оказалось возможным выработать качественные инструкции, и несогласие между разметчиками было обусловлено, по-видимому, индивидуальными особенностями восприятия. В задаче же определения тональности предложений оказалось, что наиболее сильные расхождения между разметчиками касаются нейтрального класса тональности, который необходимо изучить подробнее для того, чтобы сформулировать более точные инструкции. Также следует изучить причины достаточно сильной вариации доли предложений, тональность которых определить не удалось.

Существенной проблемой является невозможность даже в случае явно некорректных оценок разметчиков определить, были ли они вызваны несовершенством инструкций, недостаточно точной постановкой задачи или же невнимательностью или недобросовестностью самого разметчика. Например, один из разметчиков оценил тональность предложения «В книге Аствацатурова определённой темы нет, как нет и сюжета, нет, кажется, и смысла» как нейтральную, несмотря на то, что в нём прослеживается отрицательное авторское отношение к теме — книге Аствацатурова. Разметчик мог допустить такую ошибку как из-за невнимательности или пренебрежения инструкциями, так

и из-за того, что инструкции рекомендовали ориентироваться на эмоциональную составляющую предложений, а предложение не содержит яркого проявления отрицательных эмоций. Организаторам процесса разметки желательно понимать причины ошибок разметчиков, чтобы в зависимости от этого либо уточнять формулировку задачи и инструкции, либо, если кто-то из разметчиков слишком часто выносит оценки не в соответствии с инструкциями, не учитывать оценки этого разметчика. Кроме того, для построения в будущем корпусов с разметкой по тональности была бы крайне полезной возможность определять, определение тональности каких именно предложений вызывает у разметчиков наибольшие сложности, чтобы на основе этой информации целенаправленно дорабатывать инструкции. Также есть необходимость в формализованной и, желательно, не требующей большого количества ресурсов методике поиска разметчиков, не следующих инструкциям. Как показала обработка разметки предложений по наличию иронии, разметка, не соответствующая инструкциям, может достаточно сильно снижать согласие.

Можно предположить, что корпус, сформированный на основе оценок нескольких разметчиков, лучше, чем сформированный на основе оценок одного разметчика, подходит для решения практических задач, поскольку в нём «сглаживаются» особенности индивидуального восприятия. Чтобы проверить, влияет ли количество разметчиков на качество работы алгоритмов анализа тональности, был проведён следующий эксперимент. Были взяты два корпуса — 9697 предложений с разметкой каждого предложения только одним человеком (первым разметчиком, оценивавшим это предложение) и 4488 предложений для которых сошлись оценки 3–5 разметчиков. На обоих корпусах была проведена оценка качества работы рекурсивного алгоритма анализа тональности [17] и RuBERT [18]. Для рекурсивного алгоритма F_1 -мера на корпусе с разметкой только одним человеком составила 0.57; на корпусе предложений, для которых сошлись мнения всех разметчиков, — 0.69. Для RuBERT F_1 -мера на корпусе с разметкой только одним человеком составила 0.59, на корпусе предложений, для которых совпали мнения всех разметчиков — 0.72. Следовательно, число разметчиков и согласие между ними могут оказывать существенное влияние на применение корпуса для оценки качества работы алгоритмов.

Предложим рекомендации по использованию результатов разметки предложений по тональности, полученных с использованием описанной выше методики разметки. В первую очередь необходимо отбрасывать предложения, тональность которых не смог определить хотя бы один разметчик, поскольку, скорее всего, они либо неполные и не содержат законченной авторской мысли, либо их тональность невозможно было определить на основе предложенных инструкций, и разметчики, которые отметили тональность предложения как положительную, отрицательную, нейтральную или смешанную, сделали это не на основе инструкций — в любом случае для практического использования оценки тональности таких предложений бесполезны. Также стоит отбросить предложения, тональность которых хотя бы один разметчик отметил как положительную, и хотя бы один — как отрицательную, потому что если для предложения были вынесены противоположные оценки, то, скорее всего, для определения тональности этого предложения инструкции оказались бесполезными, а значит, такое предложение лучше отбросить. Если после этого согласие разметчиков окажется практически полным, т. е. доля предложений, для которых совпали оценки всех разметчиков, и альфа Криппендорфа будут выше 0.80, то в качестве тональности каждого предложения можно рассматривать оценку, данную подавляющим большинством разметчиков: 2 из 3, 3 из 4, 3 и более из 5. Если же согласие окажется хуже, то следует использовать только те предложения, для которых совпали мнения всех оценивавших их разметчиков. Нужно отметить, что при росте числа разметчиков эти рекомендации следует пересмотреть, поскольку, когда каждое предложение оценивает, например, 10 человек, вероятность того, что хотя бы один из них не сможет определить его тональность, становится достаточно высокой.

Также представляет интерес изучение вопроса о том, что такое качество разметки и что делает разметку качественной. Определить это важно, поскольку без этого невозможно понять, к какой разметке следует стремиться. В настоящий момент известна только одна черта качественной разметки — она единообразна, т. е. выполнена в соответствии с единым набором инструкций, и, как следствие, сильным согласием [19]. Остальные же свойства, или их отсутствие, требуют изучения и формулирования.

В задаче разметки предложений по тональности разметчики выставили значительно больше нейтральных оценок, чем положительных, отрицательных и смешанных. Скорее всего, причина такого дисбаланса — большое количество новостных и информационно-аналитических текстов, имеющих нейтральную тональность. Также разметчики выставили гораздо больше отрицательных оценок, чем положительных и смешанных. Можно предположить, что это обусловлено достаточно большим количеством вошедших в корпус записей из блогов и комментариев, поскольку их гораздо чаще пишут, когда автора что-то не устраивает, чем когда всё устраивает. Этот вопрос, так же, как и соотношение фразеологизмов с положительной и отрицательной тональностью, требует отдельного исследования.

Заключение

В данной работе рассмотрена организация разметки волонтерами трёх корпусов — корпуса предложений, размеченных по общей тональности на 4 класса, корпуса фразеологизмов, размеченных по тональности на 3 класса, и корпуса предложений, размеченных по наличию или отсутствию иронии. На основе имеющихся знаний о предметной области для каждой из задач были составлены инструкции для разметчиков. Также была выработана методика статистической обработки результатов разметки для оценки согласия между разметчиками. Для разметки предложений по наличию иронии показатели согласия оказались достаточно высокими, по-видимому, за счёт достаточно качественных инструкций и лёгкости самой задачи. При разметке предложений по общей тональности согласие оказалось достаточно слабым по причине сложности задачи, в частности, ввиду затруднений при выделении отличительных признаков нейтральных предложений. В то же время для 4488 предложений совпали мнения всех оценивавших их разметчиков, и такой объём корпуса представляется достаточным для решения большинства прикладных задач. При разметке фразеологизмов по тональности согласие оказалось значительно выше, чем при разметке предложений по тональности, что показывает, что предложенные инструкции для разметчиков достаточно хороши для более простых задач.

Авторы выражают благодарность д. э. н. Елене Михайловне Спиридоновой за содержательные замечания по тексту работы.

References

- [1] V. Masoumi, M. Salehi, H. Veisi, G. Haddadian, V. Ranjbar, and M. Sahebdel, *Telecrowd: A crowdsourcing approach to create informal to formal text corpora*, 2020. arXiv: [2004.11771](https://arxiv.org/abs/2004.11771) [cs.SI].
- [2] E. Mitiagina, M. Borodataya, E. Volchenkova, N. Ershova, M. Luchinina, and E. Kotelnikov, “Russian text corpus of intimate partner violence: Annotation through crowdsourcing”, in *7th International Conference on Electronic Governance and Open Society: Challenges in Eurasia. EGOSE 2020*, Springer, 2020, pp. 306–321.
- [3] S. Mohammad, “A practical guide to sentiment annotation: Challenges and solutions”, in *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, Association for Computational Linguistics, 2016, pp. 174–179.

- [4] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, “Stance and sentiment in tweets”, *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, vol. 17, no. 3, pp. 1–23, 2017.
- [5] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, and J. P. McCrae, “Corpus creation for sentiment analysis in code-mixed Tamil-English text”, in *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*, 2020, pp. 202–210.
- [6] K. Krippendorff, *Content analysis: an introduction to its methodology*. Thousand Oaks, CA: SAGE Publications, Inc., 2013.
- [7] Y. Zhao, B. Qin, and T. Liu, “Creating a fine-grained corpus for chinese sentiment analysis”, *IEEE Intelligent Systems*, vol. 30, no. 1, pp. 36–43, 2014.
- [8] J. Cohen, “A coefficient of agreement for nominal scales”, *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [9] J. Bu, L. Ren, S. Zheng, Y. Yang, J. Wang, F. Zhang, and W. Wu, “ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction”, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2021, pp. 2069–2079.
- [10] M. Navas-Loro, V. Rodríguez-Doncel, I. Santana-Perez, and A. Sánchez, “Spanish corpus for sentiment analysis towards brands”, in *Speech and Computer. SPECOM 2017*, A. Karpov, R. Potapova, and I. Mporas, Eds., Springer International Publishing, 2017, pp. 680–689.
- [11] J. L. Fleiss, “Measuring nominal scale agreement among many raters”, *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.
- [12] A. Rogers, A. Romanov, A. Rumshisky, S. Volkova, M. Gronas, and A. Gribov, “RuSentiment: An enriched sentiment analysis dataset for social media in Russian”, in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 755–763.
- [13] T. V. Zherebilo, *Slovar lingvisticheskikh terminov*. Nazran: OOO Pilgrim, 2010, in Russian.
- [14] K. Krippendorff. “Computing Krippendorff’s alpha-reliability”. (2008), [Online]. Available: https://repository.upenn.edu/asc_papers/43/ (visited on 01/17/2023).
- [15] J. Hughes, “krippendorffsalph: An R package for measuring agreement using Krippendorff’s alpha coefficient”, *The R Journal*, vol. 13, no. 1, pp. 413–425, 2021.
- [16] L. A. Jeni, J. F. Cohn, and F. De La Torre, “Facing imbalanced data—recommendations for the use of performance metrics”, in *2013 Humaine association conference on affective computing and intelligent interaction*, IEEE, 2013, pp. 245–251.
- [17] A. Y. Poletaev and I. V. Paramonov, “Recursive sentiment detection algorithm for Russian sentences”, *Modelirovanie i Analiz Informatsionnykh Sistem*, vol. 29, no. 2, pp. 134–147, 2022, in Russian.
- [18] S. Smetanin and M. Komarov, “Deep transfer learning baselines for sentiment analysis in Russian”, *Information Processing & Management*, vol. 58, no. 3, p. 102 484, 2021.
- [19] R. Artstein and M. Poesio, “Inter-coder agreement for computational linguistics”, *Computational linguistics*, vol. 34, no. 4, pp. 555–596, 2008.