Publicly available sentiment dictionary for the Russian language KartaSlovSent

Kulagin D.I.
KartaSlov.ru, Moscow, Russia
kdenisk@gmail.com

Annotation

In this paper, we describe the construction of a publicly available sentiment dictionary for the Russian language covering more than 46 thousand single words. The published dataset provides a polarity tag and a continuous polarity value from the [-1, 1] range for each of the words from a source lexicon. The paper describes a process of selecting words for the source lexicon, its part-of-speech structure, approach to data annotation using crowdsourcing, algorithm for analysis and processing of annotated data. The resulting dataset is then compared with two other sentiment dictionaries, collected independently. Dataset is available for download at https://github.com/dkulagin/kartaslov.

Keywords: sentiment lexicons, sentiment analysis, emotional words, emotive lexicon, emotional vocabulary
DOI: 10.28995/2075-7182-2021-20-1106-1119

Открытый тональный словарь русского языка КартаСловСент

Кулагин Д. И.
КартаСлов.ру, Москва, Россия
kdenisk@gmail.com

Аннотация

В статье представлен открытый тональный словарь КартаСловСент, покрывающий более 46 тысяч слов русского языка. В опубликованном датасете каждому входу в соответствие поставлена метка тональности и числовое значение силы эмоционально-оценочного заряда из непрерывного диапазона [-1, 1]. В статье описан процесс формирования словника, его частеречный состав, методика разметки входного набора при помощи краудсорсинга, приведена модель анализа и обработки полученных данных. Приведены результаты сравнения датасета КартаСловСент, включающего агрегированные данные собранной разметки и результат работы модели, с двумя лексиконами оценочной лексики, составленными независимо от представленного датасета. Словарь доступен для скачивания по адресу: https://github.com/dkulagin/kartaslov.

Ключевые слова: оценочная лексика; эмоциональная лексика; лингвистические датасеты; словари оценочной лексики; тональные словари; эмоциональный лексикон; анализ тональности; лингвистика эмоций

1 Введение

Анализ тональности текста (анализ мнений, sentiment analysis) представляет собой семейство методов прикладной лингвистики, направленных на выделение в тексте тонального компонента высказывания, включающего субъект (автора) мнения, объект, на который направлено мнение, и тональность оценки [4].

Существует три основных подхода к анализу тональности: (1) использование машинного обучения, (2) использование систем, построенных на правилах, и (3) гибридный. Преимуществом подхода, основанного на машинном обучении, является, как правило, более высокая точность [11] и способность обучаться работе в заданной предметной области. К недостаткам данного подхода можно отнести необходимость разметки обучающей выборки, что является трудоёмким и материально затратным процессом.

Системы, основанные на правилах, не требуют для своей работы предварительно размеченной обучающей коллекции текстов, но при этом используют внешний по отношению к анализируемому набору данных лингвистический ресурс – тональный словарь. Полнота и точность используемого словаря в этом случае будет влиять на качество работы системы анализа тональности [18]. Подходы, основанные на машинном обучении, и гибридные подходы могут использовать данные тонального словаря в качестве дополнительных признаков [14].

Целью настоящей работы являлось создание тонального словаря русского языка, включающего в себя как можно большее количество однословных единиц и ставящего каждой такой единице в соответствие значение силы её эмоционально-оценочного заряда из непрерывного диапазона `[-1, 1]`. Граничные значения этого диапазона соответствуют: `-1` – максимальной отрицательной оценке, `+1` – максимальной положительной оценке входной единицы. Нуль соответствует нейтральной оценке или, что одно и то же, отсутствию у входной единицы тональной окрашенности. Принцип формирования словника и максимизации покрытия однословной лексики подробно описан в соответствующем разделе статьи. Тональный словарь оформлен в виде датасета и может использоваться в составе автоматических систем анализа тональности.

Настоящая статья построена следующим образом:

- В секции «Обзор связанных работ» приведён обзор подходов к анализу тональности, дана краткая классификация тональных словарей и ссылка на обзор других словарей тональности на русском и английском языках.
- В секции «Структура датасета» описана структура опубликованного тонального словаря и приведена гиперссылка для скачивания.
- В секции «Формирование словника» описан принцип отбора входов для включения в словарь.
- В секции «Разметка» приведено описание методики разметки данных при помощи краудсорсинга.
- В секции «Модель обработки разметки» приведён анализ собранных данных и предложена модель преобразования разметки в скалярное значение силы эмоционально-оценочного заряда, а также алгоритм компиляции метки тональности.
- В секции «Сравнение с другими датасетами» приведено сравнение с лексиконами оценочных слов РуСентиЛекс (экспертная разметка) и LinisCrowd (разметка при помощи краудсорсинга).
- В секции «Заключение» приведены заключительные выводы и обозначены направления будущей работы.

2 Обзор связанных работ

Словари оценочной лексики можно классифицировать по ряду различных признаков. Существуют словари, включающие в себя только однословные языковые единицы, другие включают в себя и словосочетания. Часть словарей покрывает общую (межпредметную) лексику, другие описывают конкретную предметную область. Также можно классифицировать словари по уровню детализации шкалы тональности. Так, шкала тональности может приписывать каждому входу дискретные метки «положительное», «отрицательное», «нейтральное», или иметь более подробную дискретную градацию, или приписывать каждому входу скалярное значение из диапазона `[-1, 1]`.

Обзор существующих тональных словарей, включающий рассмотрение и совместный анализ 18 англоязычных, 8 русскоязычных словарей, а также 3 словарей, имеющих версии как для русского, так и для английского языков, приведён в [2].

Терминология, связанная с анализом тональности, не является полностью устоявшейся, следовательно, как в русскоязычных, так и в англоязычных статьях используются различные термины для обозначения одних и тех же и/или сходных понятий из рассматриваемой области. Исторический обзор работ по тональному анализу и обзор основных терминов приведён в [4].

Подходы к анализу тональности можно обобщённо разделить на три основных класса:

- Подходы, основанные на использовании машинного обучения. В этом случае алгоритм обучается на заранее размеченной по тональности коллекции текстов и в дальнейшем

может производить классификацию новых текстов. Такой подход даёт, как правило, более высокую точность [11], но требует аннотирования обучающей коллекции, что является ресурсоёмким процессом.

- Подходы, основанные на использовании правил и тональных словарей [18]. В этом случае заранее аннотированной обучающей коллекции не требуется, но необходим достаточно обширный тональный словарь. Система анализа тональности ищет в рассматриваемом тексте слова, имеющие эмоционально-оценочный заряд, и, применяя заложенные в ней правила, учитывающие отрицание и слова-усилители, вычисляет тональность всего текста.
- Гибридные подходы, совмещающие использование тональных словарей и машинного обучения.

При создании систем анализа тональности, вне зависимости от используемого подхода, необходимо учесть ряд трудных случаев, которые могут оказать влияние на качество работы системы. Среди таких случаев можно отметить следующие: одновременное наличие в тексте нескольких мнений, необходимость детектирования иронии и сарказма, учёт многозначности слов с тональной окраской и др. Обзор таких трудностей приведён в [12].

Также стоит упомянуть о связи задач анализа тональности с теорией оценочности и эмотивности. Помимо применения для целей анализа тональности, состав и структура эмоционально-оценочного лексикона исследуется в рамках лингвистики эмоций. Одним из результатов таких исследований являются лексиконы эмоционально-оценочной лексики, имеющие идеографическую организацию с иерархической структурой. Такие словари составляются вручную, вследствие чего их объём, как правило, исчисляется сотнями слов [5, 1].

3 Структура датасета

Датасет содержит только однословные языковые единицы. Размер датасета составляет 46 127 записей. Расщепление слов на отдельные значения не производится.

Входы распределены по частям речи следующим образом:

существительные	**глаголы**	**прил.-прич.**	**наречия**
47 %	32 %	19 %	2 %

Таблица 1: распределение входов по частям речи

Распределение меток тональности:

NEUT	**PSTV**	**NGTV**
60.8 %	13.5 %	25.7 %
28 049 записей	6 215 записей	11 863 записей

Таблица 2: распределение меток тональности

Распределение меток тональности по частям речи:

	NEUT	**PSTV**	**NGTV**
существительные	63.9 %	13.8 %	22.3 %
глаголы	53.8 %	12.3 %	33.9 %
прил.-прич.	64.4 %	14.7 %	20.9 %
наречия	67.6 %	12.9 %	19.5 %

Таблица 3: распределение меток тональности по частям речи

Датасет содержит следующие поля:

Поле	**Комментарий**	**Источник**
`term`	текстовый вход: слово или словосочетание	входной набор
`tag`	метка тональности из набора `<PSTV, NEUT, NGTV>`	модель
`value`	числовое значение тональности из диапазона `[-1, 1]`, где `+1` соответствует словам с максимально положительной окраской, `0` — словам с нейтральной окраской (отсутствие тональности), а `-1` — словам с максимально отрицательной окраской	модель
`pstv`	доля голосов за положительную оценку	разметка
`neut`	доля голосов за нейтральную оценку	разметка
`ngtv`	доля голосов за отрицательную оценку	разметка
`dunno`	доля голосов за ответ «Не знаю»	разметка
`pstvNgtvDisagreementRatio`	показатель рассогласованности между положительной и отрицательной оценками (см. «Уровень рассогласованности полярных оценок»)	модель

Таблица 4: описание полей датасета

Скачать тональный словарь можно по ссылке:
https://github.com/dkulagin/kartaslov/tree/master/dataset/kartaslovsent

4 Формирование словника

Формирование словника тонального словаря является отдельной нетривиальной задачей, так как составители стремятся включить в него только те языковые единицы, которые относятся к оценочной лексике или имеют эмоционально-оценочную коннотацию [3, 10]. Такой подход позволяет экономить ресурсы при разметке, сохраняя качество работы системы анализа тональности.

При формировании словника словаря КартаСловСент была поставлена задача покрыть максимальное количество однословных единиц русского языка. Таким образом, для разметки отбирались в том числе слова, не имеющие тональности (нейтральные). Ограничение было связано с методикой разметки и в словник были включены только те слова, которые являются понятными для носителей языка без дополнительных пояснений и обращения к толковому словарю. Определение понятности производилось методом опроса для слов из Викисловаря. (Данные опроса не публикуются.)

5 Разметка

Для разметки словника была использована технология краудсорсинга. Специальная программа, Лампобот, установленная на сайте kartaslov.ru, задавала пользователям сайта следующий вопрос:

- Вопрос: **`%term%`** — это что-то нейтральное, положительное или отрицательное?

Где **`%term%`** — слово или словосочетание из размечаемого набора.

Пользователям предлагалось выбрать ровно один из четырёх вариантов ответа, следующих в перечисленном ниже порядке:

- Нейтральное,
- Положительное,
- Отрицательное,
- Не знаю.

Ниже приведён пример вопроса:

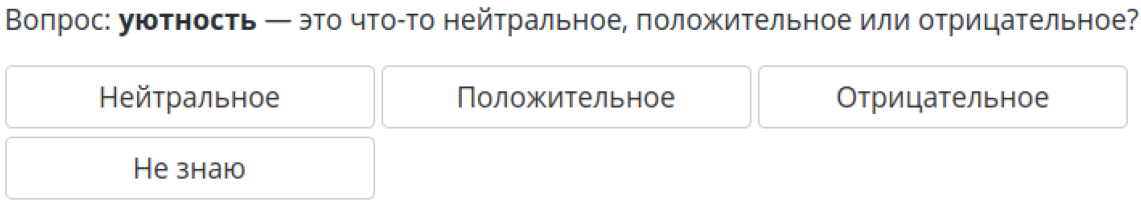

Рисунок 1: пример вопроса, задаваемого в рамках разметки при помощи краудсорсинга

При этом отвечающему не давалось дополнительного разъяснения относительно того, что подразумевается под каждым из вариантов ответа. Предполагалось, что при ответе на вопрос отвечающий задействует свою языковую интуицию и разделяемое между отвечающими понимание тональности [9, 15].

Для каждого входа было собрано не менее 25 ответов, предоставленных разными пользователями. Это число включало ответы «Не знаю». Опубликованные данные содержат долю голосов за каждый из четырёх вариантов.

6 Модель обработки разметки

Исходно разметка представляла собой набор данных, ставящий в соответствие каждому слову или словосочетанию из входного набора кортеж из четырёх элементов $(x_{neut}, x_{pstv}, x_{ngtv}, x_{dunno})$:

- x_{neut} — доля голосов за вариант «Нейтральное»;
- x_{pstv} — доля голосов за вариант «Положительное»;
- x_{ngtv} — доля голосов за вариант «Отрицательное»;
- x_{dunno} — доля голосов за вариант «Не знаю».

Для целей практического использования, а также для сравнения полученного датасета с другими тональными словарями, данные разметки были преобразованы в скалярный показатель тональности, указывающий величину силы и направление эмоционально-оценочного заряда. Затем по полученному показателю была скомпилирована метка тональности из набора <NEUT, NGTV, PSTV>.

Ниже предложена модель преобразования данных исходной разметки.

6.1 Уровень сброса

При ответе на вопрос у пользователя имелась возможность пропустить (сбросить) его и перейти к следующему. Это могло быть полезно в том случае, когда значение размечаемой языковой единицы не вполне понятно отвечающему или если однозначный ответ на вопрос программы вызывает затруднения.

Уровень сброса численно равен x_{dunno}. Ниже приведено распределение показателя x_{dunno} для всего датасета:

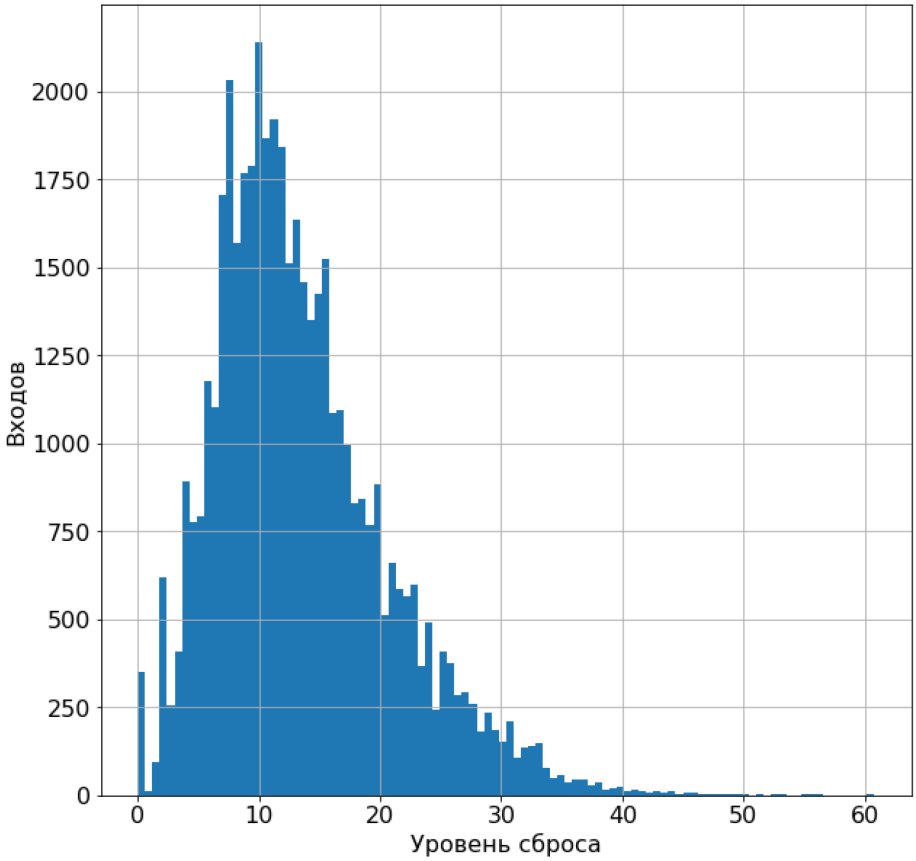

Рисунок 2: распределение значений уровня сброса по датасету

Превышение значением x_{dunno} определённого порога сигнализирует о необходимости относиться к вектору ответов по существу с долей осторожности.

6.2 Шум/сигнал

Разумно полагать, что при ответе на вопрос пользователь либо сбрасывает его и переходит к следующему, либо отвечает по существу, неявно выбирая вариант «Да, знаю». Таким образом, исходный вопрос расщепляется на каскад вопросов:

- Знаете ли вы значение **%term%** и можете ли оценить его полярность? (Варианты ответа: «Да, знаю» и «Нет, не знаю»).
- В случае ответа «Да, знаю» задаётся вопрос о полярности.

Следовательно, можно рассматривать значение x_{dunno} и вектор $(x_{neut}, x_{pstv}, x_{ngtv})$ независимо друг от друга.

Для каждого входа исходный кортеж преобразуется в тройку $(y_{neut}, y_{pstv}, y_{ngtv})$:

- y_{neut} — доля голосов за вариант «Нейтральное» (без учёта голосов за вариант «Не знаю»);
- y_{pstv} — доля голосов за вариант «Положительное» (без учёта голосов за вариант «Не знаю»);
- y_{ngtv} — доля голосов за вариант «Отрицательное» (без учёта голосов за вариант «Не знаю»).

Сумма y_{neut}, y_{pstv} и y_{ngtv} равна единице. Вектор y в общем случае раскладывается в следующую сумму:

$$y = y_{сигнал} + y_{шум}$$

- $y_{сигнал}$ отвечает верным голосам, то есть когда отвечающий понимает значение размечаемой языковой единицы и добросовестно даёт ей свою субъективную оценку;
- $y_{шум}$ отвечает остальным голосам.

Выдвинем два предположения относительно свойств разметки и векторов $y_{сигнал}$, $y_{шум}$:

- в большинстве случаев пользователи отвечали верно и ответственно, а соответственно $P(y_{сигнал}) >> P(y_{шум})$, где $P(y)$ — значение мощности, определяемое как значение максимальной компоненты вектора y;
- шум имеет бо́льшую равномерность, чем сигнал, так как ответы, вносящие вклад в сигнал, опираются на общую скрытую информацию — истинное значение тональности.

Опираясь на выдвинутые предположения, выполним следующее преобразование:

$z_{neut} = \max(0, y_{neut} - P(y) * 0.25)$

$z_{pstv} = \max(0, y_{pstv} - P(y) * 0.25)$

$z_{ngtv} = \max(0, y_{ngtv} - P(y) * 0.25)$

В рамках выдвинутых предположений такое преобразование снижает уровень шума в составе наблюдаемого значения y, но также искажает сигнал $y_{сигнал}$, усиливая в нём компонент с максимальным исходным значением.

Для иллюстрации покажем распределение итоговых значений силы эмоционально-оценочного заряда до и после преобразования. Алгоритм расчёта скалярного значения по вектору содержательных ответов приводится ниже.

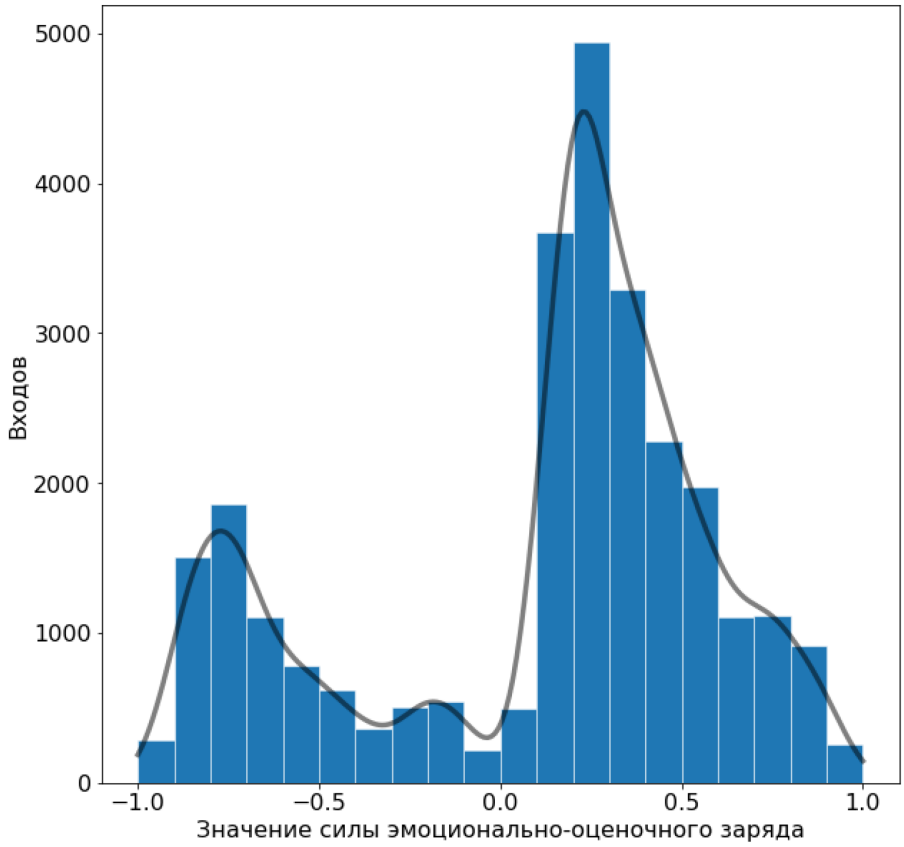

Рисунок 3: исходное распределение значений силы эмоционально-оценочного заряда (до преобразования)

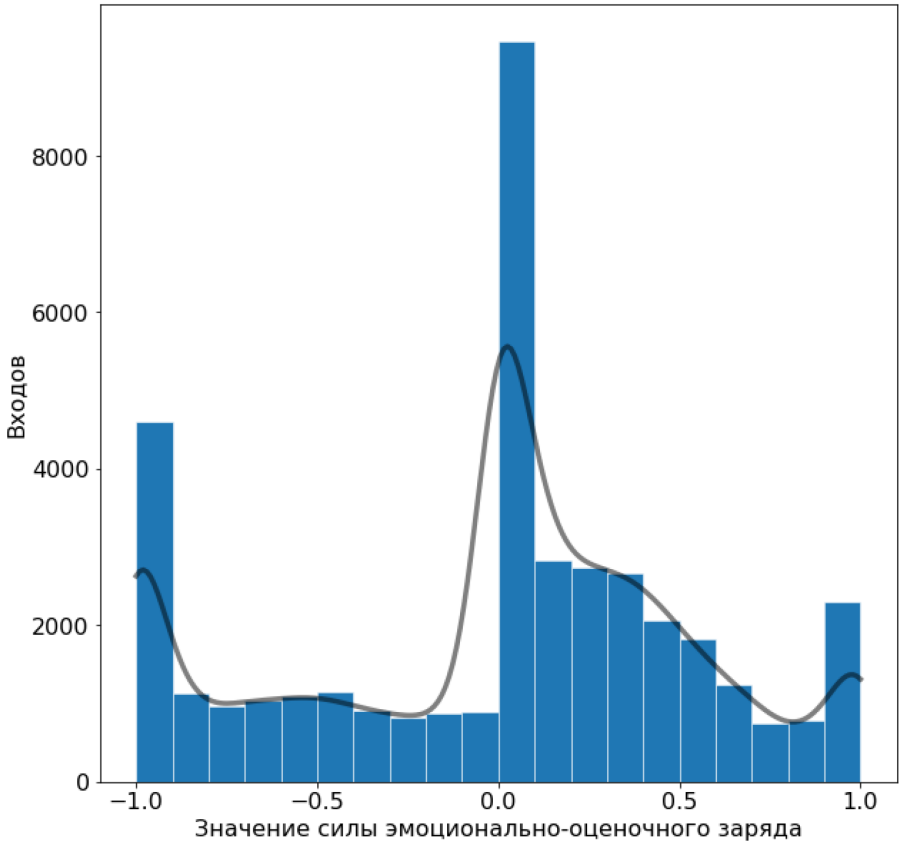

Рисунок 4: распределение значений силы эмоционально-оценочного заряда (после преобразования)

Дальнейшие действия производятся с преобразованным вектором `z`.

6.3 Числовое значение тональности

В практических приложениях оперировать вектором `z` оказывается не вполне удобно, поэтому для каждого входа было рассчитано скалярное значение силы эмоционально-оценочного заряда в диапазоне `[-1, 1]`, полюса которого отвечают максимально отрицательной и максимально положительной оценке, а нуль — нейтральной оценке входа.

При этом осмысленно выполнить такое преобразование можно не всегда. В предельном случае, когда доля голосов за нейтральный вариант равна нулю, а на полярные ответы приходится по 50 % голосов, возникает ситуация полной рассогласованности и скалярное значение силы эмоционально-оценочного заряда оказывается не определено. С другой стороны, если доля голосов за один из полярных вариантов равна нулю, то скалярное значение тональности равно доле голосов за противоположный полярный вариант.

6.4 Уровень рассогласованности полярных оценок

Уровень рассогласованности полярных оценок определяется следующим образом:

$$\texttt{pstvNgtvDisagreementRatio} = \min(z_{pstv}, z_{ngtv}) \;/\; (\max(z_{pstv}, z_{ngtv}) + z_{neut})$$

Показатель `pstvNgtvDisagreementRatio` обладает следующими свойствами:

- Лежит в диапазоне от нуля до единицы включительно.
- Равняется максимальному значению в случае, когда количество положительных ответов равно количеству отрицательных, а нейтральные ответы отсутствуют.
- Равняется нулю, когда количество положительных или отрицательных ответов равно нулю.

- Уменьшается при увеличении количества нейтральных ответов с сохранением количества голосов за полярные оценки.

Ниже приведено распределение показателя рассогласованности (для приведённых значений z):

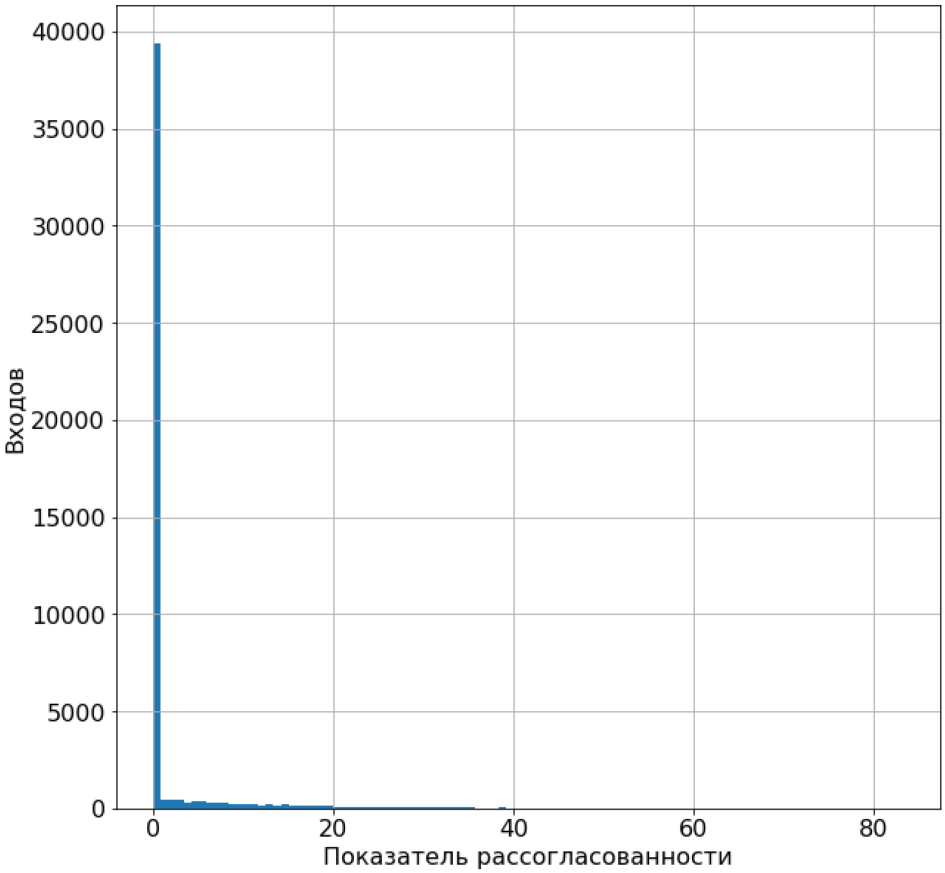

Рисунок 5: распределение значений показателя рассогласованности

Из распределения видно, что для подавляющего числа входов показатель рассогласованности полярных оценок незначителен. Приведём выборку слов с максимальными значениями показателя:

- праздность,
- антиреклама,
- утопия,
- зазноба,
- книжонка,
- блажь,
- захваливать,
- всезнайство,
- балагурство,
- сюсюканье,
- покуражиться,
- родственничек.

6.5 Алгоритм расчёта скалярного значения силы эмоционально-оценочного заряда

Определив распространённость ситуации высокой рассогласованности полярных оценок, можно предложить следующую формулу расчёта значения силы эмоционально-оценочного заряда:

$value = z_{pstv}$, если $z_{pstv} \geq z_{ngtv}$

$value = -z_{ngtv}$, если $z_{pstv} < z_{ngtv}$

В этой формуле голоса за полярную оценку с меньшей долей добавляются к голосам за ответ «Нейтральное». До упрощения формула выглядит следующим образом:

$value = +1 * z_{pstv} + 0 * (z_{neut} + z_{ngtv})$, если $z_{pstv} \geq z_{ngtv}$

$value = -1 * z_{ngtv} + 0 * (z_{neut} + z_{pstv})$, если $z_{pstv} < z_{ngtv}$

Доверие этому значению тем выше, чем ниже показатель рассогласованности полярных оценок.

6.6 Компиляция метки тональности

Граница между полярными и нейтральными языковыми единицами не является точно определённой, следовательно, как и в других языковых подсистемах, существуют более хорошие и менее хорошие представители тональных классов [16, 17].

Выбор границ между тональными классами зависит от решаемой задачи и выполняется пользователем датасета. Ниже приводится вариант «по умолчанию», реализованный для компиляции метки в опубликованном наборе данных. Далее, в разделе сравнения с другими датасетами, предлагается конструктивный способ выбора пороговых значений тональных классов, опирающийся на данные калибровочного словаря.

tag = PSTV, если value ≥ 0.55

tag = NGTV, если value ≤ -0.35

tag = NEUT, если value $\in (-0.35, 0.55)$

Прим. С учётом соображений, описанных выше, более разумной выглядит пятиступенчатая система меток тональности <NGTV, NGTV/NEUT, NEUT, NEUT/PSTV, PSTV>. В такой системе существуют два буферных класса NGTV/NEUT и NEUT/PSTV. В прикладной задаче можно в таком случае полагать, что языковая единица, принадлежащая данному буферному классу, одновременно принадлежит и полярному, и нейтральному классу. Скажем, при тесте на эквивалентность единица из класса NGTV будет соотноситься как с другими единицами из класса NGTV, так и с другими единицами из класса NGTV/NEUT.

7 Сравнение с другими датасетами

Для целей сравнения были взяты два оригинальных (непереводных) датасета оценочных слов русского языка: РуСентиЛекс [3] и LinisCrowd [10]. Отметим, что РуСентиЛекс размечался экспертами вручную, в то время как для создания лексикона LinisCrowd использовался краудсорсинг.

7.1 Сравнение с лексиконом РуСентиЛекс

Присвоение тональных меток входам лексикона РуСентиЛекс производилось экспертом. Для целей сравнения словарь РуСентиЛекс был отфильтрован следующим образом:

- были оставлены только записи, промаркированные как «positive», «negative» и «neutral» («positive/negative» исключаются из рассмотрения);
- были оставлены только уникальные записи, для которых нет расщепления на разные лексические значения, имеющие различную тональность;
- все входы, у которых есть хотя бы одна запись с меткой «neutral», были добавлены в результирующий набор с меткой «Нейтральное».

Выполнение последнего пункта гарантировало наполнение класса «Нейтральное» в датасете РуСентиЛекс, который в противном случае оказывался слишком маленьким, чем нарушалась стабильность процесса калибровки.

Датасет КартаСловСент был отфильтрован следующим образом:

- уровень сброса не более 25 %;
- уровень рассогласованности полярных оценок не более 10 %.

Ниже приведена характеристика общего лексикона обоих словарей, участвующих в сравнении:

	КартаСловСент (после калибровки)	**РуСентиЛекс**
«Нейтральное»	772 слова	853 слова
«Положительное»	1 774 слова	1 654 слова
«Отрицательное»	4 316 слов	4 355 слов
Всего: 6 862 слова		

Таблица 5: характеристика общего лексикона словарей КартаСловСент и РуСентиЛекс

7.2 Калибровка порогов тональных меток

Так как исходно в модели имеется возможность оперировать скалярным значением силы эмоционально-оценочного заряда из непрерывного диапазона `[-1, 1]`, а метки компилируются на основе алгоритма, то перед сравнением двух наборов данных был выполнен поиск оптимальных пороговых значений, которые обеспечивали наибольшее пересечение. Для этого алгоритм компиляции метки тональности был параметризован следующим образом:

$tag = PSTV$, если $value \geq t_{pstv}$

$tag = NGTV$, если $value \leq t_{ngtv}$

$tag = NEUT$, если $value \in (t_{ngtv}, t_{pstv})$

Мера совпадения α_{tag} вычислялась отдельно по каждому тональному классу. Затем максимизировалась сумма показателей совпадения по всем трём классам α:

$\alpha = \alpha_{neut} + \alpha_{pstv} + \alpha_{ngtv}$

Пороги t_{pstv} и t_{ngtv} подбирались таким образом, чтобы максимизировать значение α.

В качестве меры совпадения α_{tag} по каждому из тональных классов был выбран аналог `F1`-меры:

$\alpha_{tag} = |A_{tag} \cap B_{tag}| \ / \ (|A_{tag} \cap B_{tag}| + ½ * (|A_{tag} \setminus B_{tag}| + |B_{tag} \setminus A_{tag}|))$

Где A_{tag} — множество входов с меткой `tag` (`NEUT`, `PSTV` или `NGTV`) в первом датасете, B_{tag} — во втором. Отметим, что такой показатель не зависит от порядка сравнения.

7.3 Результаты сравнения с лексиконом РуСентиЛекс

Результаты калибровки порогов и результаты сравнения приведены ниже:

t_{ngtv}	t_{pstv}	α_{neut}	α_{pstv}	α_{ngtv}	**Процент совпадений по всем входам**
-0.45	0.47	47.8 %	89.3 %	93 %	**87 %**

Таблица 6: результаты сравнения словарей КартаСловСент и РуСентиЛекс

Разнополярные метки были присвоены 44 словам (в скобках приведён источник тональности согласно датасету РуСентиЛекс):

КартаСловСент = PSTV, РуСентиЛекс = NGTV	**КартаСловСент = NGTV, РуСентиЛекс = PSTV**
• лицеприятный (opinion) • красивость (opinion) • великодержавный (opinion) • лечение (fact)	• пафос (feeling) • бедненький (opinion) • безыскусность (opinion) • дотошность (opinion)

• заботить (feeling) • мудрствовать (opinion) • мудрствование (opinion) • угорать (fact) • крепкоголовый (opinion) • ностальгический (feeling) • соскучиться (feeling) • фаворитка (fact) • любительство (opinion) • невозмутимый (opinion) • сотворить (opinion) • неотступность (opinion) • ностальгия (feeling) • наставительность (opinion) • озаботиться (feeling) • миньон (opinion) • переплюнуть (opinion) • неотступный (opinion) • идеализировать (opinion) • засмущаться (feeling)	• дотошный (opinion) • безыскусный (opinion) • зашибиться (feeling) • стебать (opinion) • лукавый (opinion) • лукавость (opinion) • коварство (opinion) • коварный (opinion) • вероломство (opinion) • вероломный (opinion)

Таблица 7: примеры входов с разнополярными метками согласно словарям КартаСловСент и РуСентиЛекс

7.4 Сравнение с лексиконом LinisCrowd

Лексикон LinisCrowd, также как и датасет КартаСловСент, размечался при помощи краудсорсинга. Исходные данные лексикона представляли собой голоса из набора <-2, -1, 0, 1, 2>. Для целей сравнения была проведена взаимная калибровка двух датасетов с подбором границ классов в каждом наборе данных.

Скалярная оценка в лексиконе LinisCrowd рассчитывалась по следующей формуле:

$$value_{linis} = (-1 * c_{-2} + (-0.5) * c_{-1} + 0 * c_0 + 0.5 * c_1 + 1 * c_2) / (c_{-2} + c_{-1} + c_0 + c_1 + c_2)$$

где c_a — количество голосов за оценку a из набора <-2, -1, 0, 1, 2>.

Компиляция тональной метки для датасета LinisCrowd была параметризована аналогичным образом:

$$tag_{linis} = PSTV, \text{ если } value_{linis} \geq t_{linis\_pstv}$$
$$tag_{linis} = NGTV, \text{ если } value_{linis} \leq t_{linis\_ngtv}$$
$$tag_{linis} = NEUT, \text{ если } value_{linis} \in (t_{linis\_ngtv}, t_{linis\_pstv})$$

Затем была проведена взаимная калибровка порогов.

Ниже приведена характеристика общего лексикона словарей, участвующих в сравнении:

	КартаСловСент (после калибровки)	**LinisCrowd (после калибровки)**
«Нейтральное»	2 616 слов	2 713 слов
«Положительное»	1 085 слов	1 055 слов
«Отрицательное»	1 633 слова	1 566 слов
Всего: 5 334 слова		

Таблица 8: характеристика общего лексикона словарей КартаСловСент и LinisCrowd

7.5 Результаты сравнения с лексиконом LinisCrowd

Результаты калибровки порогов и результаты сравнения приведены ниже:

t_{ngtv} (КартаСловСент)	t_{pstv} (КартаСловСент)	t_{linis_ngtv} (LinisCrowd)	t_{linis_pstv} (LinisCrowd)
-0.47	0.62	-0.21	0.09

Таблица 9: результаты взаимной калибровки порогов для словарей КартаСловСент и LinisCrowd

α_{neut}	α_{pstv}	α_{ngtv}	**Процент совпадений по всем входам**
78.6 %	68.3 %	82.8 %	**77.8 %**

Таблица 10: результаты сравнения словарей КартаСловСент и LinisCrowd

Разнополярные метки были присвоены 44 словам (это не опечатка — количество совпало для датасетов LinisCrowd и РуСентиЛекс).

8 Заключение

Два независимых эксперимента, направленных на сравнение тонального словаря КартаСловСент со словарями РуСентиЛекс и LinisCrowd в области пересечения лексиконов, показали высокий процент совпадений тональных меток. Следует отметить, что со стороны КартаСловСент сравнивались дискретные тональные метки, полученные в результате применения модели к исходной разметке.

В численном выражении словарь КартаСловСент предлагает наибольшее покрытие однословной лексики среди оригинальных (непереводных) тональных словарей русского языка, согласно данным сравнительного обзора тональных словарей [2]. При этом стоит отметить, что приведённые в обзоре словари могут включать в себя однословные лексические единицы, не представленные в словаре КартаСловСент, а также многословные языковые единицы.

Отличительным свойством словаря КартаСловСент является поддержка непрерывной шкалы для определения значения силы эмоционально-оценочного заряда, не представленной в других словарях, приведённых в обзоре.

Требуют дополнительного исследования некоторые детали в модели обработки получаемой разметки. В том числе необходимо изучить влияние выбираемого значения для порога отсечения шума на результаты вычисления числового значения тональности.

Возможным направлением исследований является проведение сравнительного эксперимента по оценке тональности текстов с использованием различных словарей тональности русского языка.

Одним из направлений развития датасета является указание мотивации наличия у языковых единиц эмоционально-оценочного заряда, а для случая эмоциональной мотивации — указание ассоциации с основными эмоциями [6, 7, 13, 19]. Полезной для пользователей датасета может оказаться возможность узнать источник тональности [8, 3].

Другим направлением развития является расширение лексического покрытия словаря и включение в него многословных языковых единиц.

Библиография

[1] Бабенко Л. Г. Лексические средства обозначения эмоций в русском языке. – Изд-во Уральского университета, 1989.

[2] Котельников Е. В. и др. Современные словари оценочной лексики для анализа мнений на русском и английском языках (аналитический обзор) //Научно-техническая информация. Серия 2: Информационные процессы и системы. – 2020. – №. 12. – С. 16-33.

[3] Лукашевич Н. В., Левчик А. В. Создание лексикона оценочных слов русского языка РуСентилекс //Открытые семантические технологии проектирования интеллектуальных систем. – 2016. – №. 6. – С. 377-382.
[4] Семина Т. А. Анализ тональности текста: современные подходы и существующие проблемы //Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 6, Языкознание: Реферативный журнал. – 2020. – №. 4.
[5] Фомина З. Е. Немецкая эмоциональная картина мира и лексические средства её вербализации. – 2006.
[6] Cambria E. et al. Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems //Development of multimodal interfaces: active listening and synchrony. – Springer, Berlin, Heidelberg, 2010. – С. 148-156.
[7] Cambria E., Livingstone A., Hussain A. The hourglass of emotions //Cognitive behavioural systems. – Springer, Berlin, Heidelberg, 2012. – С. 144-157.
[8] Feng S. et al. Connotation lexicon: A dash of sentiment beneath the surface meaning //Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2013. – С. 1774-1784.
[9] Hu M., Liu B. Mining and summarizing customer reviews //Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. – 2004. – С. 168-177.
[10] Koltsova O. Y., Alexeeva S., Kolcov S. An opinion word lexicon and a training dataset for Russian sentiment analysis of social media //Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE. – 2016. – Т. 2016. – С. 277-287.
[11] Kotelnikova A. V. Comparison of Deep Learning and Rule-based Method for the Sentiment Analysis Task //2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon). – IEEE, 2020. – С. 1-6.
[12] Loukachevitch N. Automatic Sentiment Analysis of Texts: The Case of Russian //The Palgrave Handbook of Digital Russia Studies. – Palgrave Macmillan, Cham, 2021. – С. 501-516.
[13] Mohammad S. M., Turney P. D. Crowdsourcing a word–emotion association lexicon //Computational intelligence. – 2013. – Т. 29. – №. 3. – С. 436-465.
[14] Mohammad S. M., Kiritchenko S., Zhu X. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets //arXiv preprint arXiv:1308.6242. – 2013.
[15] Mohammad S. A practical guide to sentiment annotation: Challenges and solutions //Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis. – 2016. – С. 174-179.
[16] Nöth W. Symmetries and Asymmetries between Positive and Negative Emotion words // Proceedings of the Conference of the German Association of University Professors of English / Ed. by Wilhelm L. Busse. – Vol. XIII. – Tübungen: Max Niemeyer, 1992. – PP. 72-88.
[17] Rosch E. H. Natural categories //Cognitive psychology. – 1973. – Т. 4. – №. 3. – С. 328-350.
[18] Taboada M. et al. Lexicon-based methods for sentiment analysis //Computational linguistics. – 2011. – Т. 37. – №. 2. – С. 267-307.
[19] Zaśko-Zielińska M., Piasecki M., Szpakowicz S. A large wordnet-based sentiment lexicon for Polish //Proceedings of the International Conference Recent Advances in Natural Language Processing. – 2015. – С. 721-730.