MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 30, NO. 4, 2023
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Semantic rule-based sentiment detection algorithm for Russian publicism sentences

A. Y. Poletaev[1], I. V. Paramonov[1], E. I. Boychuk[1]

DOI: 10.18255/1818-1015-2023-4-394-417

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

MSC2020: 68T50
Research article
Full text in Russian

Received November 6, 2023
After revision November 24, 2023
Accepted November 29, 2023

The article is devoted to the task of sentiment detecton of Russian sentences, which is understood as the author's attitude on the sentence topic expressed through linguistic expression features. Today most studies on this subject utilize texts of colloquial style, limiting the applicability of their results to other styles of speech, particularly to the publicism.
To fill the gap, the authors developed a novel publisism sentences oriented sentiment detection algorithm. The algorithm recursively applies appropriate rules to sentence parts represented as constituency trees. Most of the rules were proposed by a philology expert, based on knowledge on the expression features from Russian philology, and algorithmized using constituency trees generated by the algorithm. A decision tree and a sentiment vocabulary are also used in the work. The article contains the results of evaluation of the algorithm on the publicism sentences corpus OpenSentimentCorpus, F-measure is 0.80. The results of errors analysis are also presented.

Keywords: sentiment analysis; sentiment detection; semantic rules; publicism; constituency tree

INFORMATION ABOUT THE AUTHORS

Anatoliy Y. Poletaev corresponding author	orcid.org/0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru Post-graduate student.
Ilya V. Paramonov	orcid.org/0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org PhD, Associate professor.
Elena I. Boychuk	orcid.org/0000-0001-6600-2971. E-mail: elena-boychouk@rambler.ru Doctor of Science, Associate professor.

Funding: The reported study was funded by the grant of Russian Science Foundation No. 23-21-00495.

For citation: A. Y. Poletaev, I. V. Paramonov, and E. I. Boychuk, "Semantic rule-based sentiment detection algorithm for Russian publicism sentences", *Modeling and analysis of information systems*, vol. 30, no. 4, pp. 394-417, 2023.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 30, № 4, 2023
сайт журнала: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Алгоритм определения тональности предложений публицистического стиля на русском языке на основе семантических правил

А. Ю. Полетаев[1], И. В. Парамонов[1], Е. И. Бойчук[1]

DOI: 10.18255/1818-1015-2023-4-394-417

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

УДК 004.912+10.02.21
Научная статья
Полный текст на русском языке

Получена 6 ноября 2023 г.
После доработки 24 ноября 2023 г.
Принята к публикации 29 ноября 2023 г.

Статья посвящена задаче определения тональности предложения на русском языке, понимаемой как отношение автора предложения к его теме, выраженное с помощью языковых средств. В настоящий момент большинство исследований по этой теме проводятся на текстах разговорного стиля речи, что ограничивает применимость их результатов для других стилей, в частности, публицистического.
Для того, чтобы заполнить этот пробел, авторами был разработан алгоритм определения тональности, ориентированный на применение к предложениям публицистического стиля речи. Алгоритм рекурсивно применяет подходящие правила к составным частям предложения, представленным в виде дерева синтаксических единиц. Большинство правил было построено на основе знаний эксперта-филолога относительно средств выражения тональности, известных русской лингвистике, и выбора тех из них, которые достаточно формализованы для того, чтобы их можно было алгоритмизировать с использованием генерируемых в рамках алгоритма деревьев синтаксических единиц. Также применялись дерево решений и тональный словарь. В статье приведены результаты эксперимента по апробации предложенного алгоритма на корпусе предложений публицистического стиля OpenSentimentCorpus, F-мера составила 0.80, а также результаты анализа ошибок алгоритма.

Ключевые слова: анализ тональности; определение тональности; семантические правила; публицистический стиль; дерево синтаксических единиц

ИНФОРМАЦИЯ ОБ АВТОРАХ

Анатолий Юрьевич Полетаев автор для корреспонденции	orcid.org/0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru аспирант.
Илья Вячеславович Парамонов	orcid.org/0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org кандидат физико-математических наук, доцент.
Елена Игоревна Бойчук	orcid.org/0000-0001-6600-2971. E-mail: elena-boychouk@rambler.ru доктор филологических наук, доцент.

Финансирование: Исследование выполнено за счет гранта Российского научного фонда № 23-21-00495.

Для цитирования: A. Y. Poletaev, I. V. Paramonov, and E. I. Boychuk, “Semantic rule-based sentiment detection algorithm for Russian publicism sentences”, *Modeling and analysis of information systems*, vol. 30, no. 4, pp. 394-417, 2023.

Введение

Автоматический анализ тональности предложений — одна из важных и популярных задач компьютерной лингвистики, заключающаяся в определении отношения автора предложения к его теме [1]. В ходе автоматического анализа тональности конкретному предложению сопоставляется один из 2–4 классов тональности (положительный, отрицательный, нейтральный, смешанный).

Подавляющее большинство методов определения тональности предложений используют либо машинное обучение, либо сформулированные экспертами семантические правила [2—4], причём первая категория методов преобладает. Методы, основанные на семантических правилах, крайне немногочисленны даже для английского языка [5, 6] и практически полностью отсутствуют для русского.

Кроме того, бóльшая часть методов разработана для анализа предложений, извлечённых из записей в социальных сетях или интернет-отзывов на товары и услуги, относящихся преимущественно к разговорному стилю речи [7, 8]. В то же время, алгоритмы, достаточно хорошо работающие с предложениями разговорного стиля, значительно менее качественно определяют тональность предложений, относящихся к другим стилям речи. В частности, в предыдущей работе авторов [9] F_1-мера для рекурсивного алгоритма выведения тональности на корпусе интернет-отзывов на отели составила 0.75, тогда как на корпусе публицистических текстов OpenCorpora — лишь 0.70.

Такая разница в результатах, скорее всего, обусловлена тем, что для выражения авторской позиции в публицистическом стиле используется иной спектр речевых средств, чем в разговорном. Эта гипотеза подтверждается лингвистическими исследованиями, осуществленными на материале отзывов и публицистических текстов. Так, для определения тональности в отзывах лингвистами были выделены такие речевые средства, как частотное употребление специфических для данного жанра лексем, в особенности с положительной или отрицательной эмоционально-экспрессивной окраской, сочетаемость слов в предложении, роль отрицания, грамматических форм глаголов, употребление количественных наречий, прилагательных в превосходной степени [7, 10], также выделяются конкретные средства выражения положительной, отрицательной и нейтральной тональностей в интернет-отзывах [8]. Что касается публицистического текста, то его характер априори предполагает меньшую степень эмотивности, что может вести за собой проблему имплицитности проявления тональных средств, а также изменение их характера и перечня в целом [11]. Ввиду этого алгоритм определения тональности, хорошо работающий со средствами, активно используемыми в разговорном стиле, не всегда будет так же хорошо работать со средствами публицистического стиля. Данные соображения, в совокупности с общей неисследовательностью вопросов, связанных с алгоритмами анализа тональности русскоязычных текстов, основанными на семантических правилах, и определили направленность настоящей работы.

Цель данной статьи — разработать алгоритм определения тональности русскоязычного предложения, основанный на семантических правилах и ориентированный на применение к текстам в публицистическом стиле в широком смысле (новости, статьи в СМИ, записи блогов и т. п.), а также оценить его эффективность. Тональность понимается как отношение автора предложения к его теме, выраженное с помощью языковых средств. При разработке алгоритма использовались идеи рекурсивного алгоритма выведения тональности, описанного в предыдущей работе авторов [9], при этом заключает в себе более глубокий анализ языковых средств, используемых для выражения авторского отношения, в том числе в публицистическом стиле речи. Для оценки эффективности используется OpenSentimentCorpus, построенный с использованием краудсорсинга на основе корпуса публицистических текстов OpenCorpora [12].

Оставшаяся часть работы организована следующим образом. В разделе 1 приводится информация о лингвистических ресурсах, используемых в работе. В разделе 2 описывается предлагаемый

алгоритм. Раздел 3 содержит результаты экспериментов по оценке качества определения тональности алгоритмом, в том числе анализ ошибок. В заключении подведены итоги работы.

1. Лингвистические ресурсы

1.1. Корпус

Для экспериментов в работе используется OpenSentimentCorpus — корпус предложений на русском языке, извлечённых из новостных и публицистических текстов, относящихся к различным предметным областям, с разметкой по тональности [12]. Каждое предложение, входящее в OpenSentimentCorpus, оценивалось как минимум тремя разметчиками. Разметчики могли отметить предложение как имеющее положительную, отрицательную или смешанную (положительную и отрицательную) тональность, как нейтральное или как сомнительное, если его смысл было определить невозможно. В OpenSentimentCorpus вошли те предложения, для которых подавляющее большинство оценивавших их разметчиков сошлись во мнениях, и ни один разметчик не оценил их как сомнительное. Таким образом, в корпус были включены только предложения, мнение о тональности которых достаточно единообразно, и эксперименты, проводимые на такой разметке, можно считать достаточно показательными. Так как в данной работе решается задача классификации на три класса тональности, в ходе экспериментов использовались только предложения из OpenSentimentCorpus с положительной, отрицательной или нейтральной тональностью.

Поскольку точность исходной разметки для лингвистических методов является существенной, перед началом работы над алгоритмом OpenSentimentCorpus был выборочно перепроверен экспертом-лингвистом. Для перепроверки было случайно выбрано 240 предложений (по 80 каждого класса), разметка которых проверялась на соответствие используемому определению тональности. Было обнаружено, что 18 из проверенных экспертом предложений были размечены неверно, для них разметка была исправлена. Большая часть встреченных ошибок относилась к предложениям, содержащим косвенное цитирование, и ошибочно принятым разметчиками за тональные, например: «В ряде случаев, по её мнению, законопроекты вступают в противоречие с действующим законодательством и создают условия для ограничения конституционных прав и свобод граждан» или «Женщина считает, что руководство клиники, таким образом, нарушило её конституционное право на свободу вероисповедания».

Всего OpenSentimentCorpus содержит 4487 предложений, из которых 533 имеют положительную тональность, 1495 — отрицательную, а оставшиеся 2459 нейтральны.

1.2. Словари тональных слов

Для русского языка доступны два достаточно развитых словаря тональных слов: RuSentiLex-2017 [13] и KartaSlovSent [14]. Поскольку для анализа тональности крайне важно качество используемого тонального словаря, была проведена выборочная экспертная оценка этих словарей. В ходе экспертной оценки из каждого словаря было выбрано по 250 слов каждого из классов тональности. Эксперты определяли, действительно ли те слова, которые обозначены в словаре как имеющие положительную или отрицательную тональность, являются признаками выражения автором эмоции или вынесения оценок вне зависимости от контекста, в котором они употреблены. Также в ходе оценки определялось, насколько сильно в словарях пересекаются наборы тональных слов.

Было выявлено, что набор одиночных тональных слов в KartaSlovSent гораздо шире, чем в RuSentiLex-2017, но в RuSentiLex также есть достаточно много тональных слов, отсутствующих в KartaSlovSent. Также был обнаружен существенный недостаток KartaSlovSent: достаточно большому числу нейтральных слов, которые могут использоваться как в положительном, так и в отрицательном контексте, присвоены положительные метки тональности, например, глаголам «уметь», «основать», «изучить», существительному «сила». Это может быть вызвано тем, что разметчики, видевшие эти слова вне контекста, считали, что чаще всего эти слова встречаются в положи-

тельном контексте, и присваивали им положительную метку. Однако поскольку сами по себе эти слова нейтральны и, более того, могут встречаться в языке в фразах совершенно разной тональности, например, «сила гравитации», «сила благотворительности» и «сила коррупции» — включение их в словарь тональных слов, используемый алгоритмом, основанным на правилах, будет приводить к ошибкам определения тональности предложения. В то же время экспертная оценка показала, что состав слов, которым в KartaSlovSent назначена отрицательная метка тональности, достаточно точен.

2. Алгоритм определения тональности

В данной работе используются наработки рекурсивного алгоритма выведения тональности предложений на русском языке, основанного на использовании дерева синтаксических единиц [9]. Этот алгоритм рассматривается в качестве исходного. Данная работа содержит ряд существенных улучшений исходного алгоритма, основанных на более тонком анализе механизмов выражения тональности в русском языке, характерных в том числе для предложений публицистического стиля.

2.1. Общая схема работы

Алгоритм определяет тональность фразы (узла дерева синтаксических единиц) по тональностям её частей (потомков этого узла) с помощью набора правил, описывающего, как именно в языке выражается тональность. Тональности отдельных слов (листьев дерева синтаксических единиц), а также устойчивых словосочетаний, определяется с помощью тонального словаря. Опишем схему работы алгоритма более формально. Пусть N — синтаксическая единица (фраза), $C(N)$ — множество дочерних синтаксических единиц, из которых состоит N, а $S(N)$ — тональность N. Алгоритм начинает свою работу от корня дерева синтаксических единиц N_r, представляющего собой всё предложение; $S(N_r)$ — искомая тональность предложения. Тональность $S(N)$ для заданного N вычисляется следующим образом:

- если текст N присутствует в тональном словаре, то $S(N)$ — словарная тональность N;
- иначе, если $C(N) = \emptyset$, т. е. фраза состоит из одного слова, отсутствующего в тональном словаре, то $S(N)$ — нейтральная тональность;
- иначе рекурсивно вычислить $S(N_c)$ для каждого $N_c \in C(N)$, выбрать подходящее правило и вычислить по нему $S(N)$.

В зависимости от того, как тональность фразы определяется по тональностям её частей, было выделено две группы правил.

В первой группе тональность фразы является результатом простого соединения тональностей своих частей. Например, нейтральное подлежащее «законопроект», соединяясь с положительным определением «своевременный», делает тональность фразы «своевременный законопроект» положительной. Результат соединения может быть различным в зависимости от состава фразы. Создание этой группы правил рассмотрено в подразделе 2.4.

Вторая группа правил описывает определение тональности фраз, содержащих специальные языковые средства выражения тональности, то есть таких, тональность которых не является результатом простого соединения тональностей их частей. Наиболее большую группу среди них составляют фразы с отрицаниями, разработка правила для определения тональности которых рассмотрена в подразделе 2.2. Остальные правила описаны в подразделе 2.5.

Ещё одна группа относится к фразам, состоящим из однородных членов предложения, то есть представленным синтаксическими единицами типов однородные-подлежащие, однородные-сказуемые и т. п., например, «ум, честь и совесть» (положительная тональность), «разработанный коллективно и обречённый на неудачу» (отрицательная тональность). Для данных случаев было составлено следующее правило определения тональности: фраза имеет положительную тональность, если хотя бы одна из её однородных частей имеет положительную тональность, и тональность

Table 1. Baseline sentiment classification performance

Таблица 1. Качество работы исходного алгоритма

Класс предложений	Точность	Полнота	F_1-мера	Количество предложений
Положительный	0.46	0.39	0.43	533
Нейтральный	0.76	0.76	0.76	2459
Отрицательный	0.69	0.72	0.71	1495
Среднее	0.64	0.62	0.63	4487
Взвешенное среднее	0.70	0.70	0.70	4487

Доля правильных ответов алгоритма (accuracy) = 0.70

Table 2. Baseline sentiment classification confusion matrix

Таблица 2. Матрица ошибок исходного алгоритма

предсказ. / реальн.	Положит.	Нейтр.	Отрицат.	Всего
Положительная	209	249	75	533
Нейтральная	179	1873	407	2459
Отрицательная	62	355	1078	1495

ни одной из частей не отрицательна; фраза имеет отрицательную тональность, если хотя бы одна из её однородных частей имеет отрицательную тональность, и тональность ни одной из частей не положительна; иначе фраза нейтральна.

В данной работе используется построитель деревьев синтаксических единиц, описанный в [15], использующий более точную, по сравнению с [9], грамматику — в частности, он различает прямые и косвенные дополнения, а также различные типы придаточных предложений. Показатели качества работы исходного алгоритма приведены в таблицах 1 и 2. Как можно видеть, такой алгоритм достаточно неплохо отделяет предложения с отрицательной тональностью от нейтральных, однако с трудом обнаруживает положительную тональность.

2.2. Правило для фраз с отрицаниями

В рамках данного исследования фраза с отрицанием — синтаксическая единица, состоящая из двух частей, одна из которых — отрицание «не», «нет», «ни один»; например, «не ошибаться» или «не был удачным» (более сложные варианты фраз с отрицанием не рассматриваются). Наиболее сложный вопрос обработки таких фраз — определение тональности фразы, отрицаемая часть которой нейтральна. Если отрицание фразы с положительной тональностью, как правило, имеет отрицательную тональность, а отрицание фразы с отрицательной тональностью — положительную, то отрицание нейтральной информации может иметь отрицательную тональность, например, «Объяснений от Павла мы не получили», либо быть нейтральным, например, «Пик работы АСВ не пройден».

Для того, чтобы построить правило для определения тональности фраз с отрицаниями, из корпуса были выбраны все уникальные фразы-отрицания (синтаксические единицы). Всего таких фраз оказалось 861. Каждая фраза была вручную размечена по тональности двумя экспертами, и для 239 фраз их разметка совпала. Эта разметка была принята как близкая к объективной. Поскольку, как показано в таблице 3, в ней оказалось достаточно много как отрицательных, так и нейтральных отметок, дальнейшая работа проводилась именно на этих 239 фразах.

В первую очередь, на рассматриваемых фразах были проверены два правила для обработки отрицаний, использующие только информацию о тональности отрицаемой части: первое считало тональность фразы с отрицаемой нейтральной частью отрицательной (т. е. было аналогично тому, которое использовалось в работе [9]) и показало точность определения итоговой тональности фразы 0.51; второе считало тональность фразы с отрицаемой нейтральной частью отрицательной и показало точность 0.46. Поскольку такой точности явно недостаточно для использования этих правил

Table 3. Distribution of sentiment marks of negated parts of phrases with negations and entire phrases with negations

Таблица 3. Распределение оценок тональности отрицаемой части фраз с отрицаниями и фраз с отрицаниями целиком

Тональность	Положительная	Нейтральная	Отрицательная
отрицаемой части	25	190	24
фразы целиком	38	84	117

в алгоритме, было принято решение использовать для более точного выведения тональности фраз с отрицаниями дополнительную информацию.

Для построения более точного алгоритма для каждой из 239 фраз была собрана следующая информация:

- тональность отрицаемой части,
- тип синтаксической единицы отрицаемой части,
- частеречная разметка (PoS-теги) всех слов в отрицаемой части,
- время всех глаголов отрицаемой части, если они в ней присутствуют,
- лицо всех слов отрицаемой части, которые изменяются по лицу.

Для того, чтобы определить, по каким из имеющихся признаков можно наиболее точно определить тональность фразы в целом, на этих данных было построено дерево решений [16]. Дерево решений строится итеративно, на каждом шаге среди всех возможных критериев присвоения наблюдению класса (т. е. определения тональности фразы с отрицанием) выбирается тот, с помощью которого можно наиболее точно определить класс наибольшего числа наблюдений среди тех, которым на предыдущих шагах класс присвоен ещё не был. Процесс останавливается, когда достигается максимальное число критериев, предварительно выбираемое вспомогательным алгоритмом кросс-валидации так, чтобы минимизировать количество неверно классифицированных наблюдений на каждой из валидационных выборок. Каждому из критериев дерева решений была дана лингвистическая интерпретация, и на их основе был построен следующий алгоритм.

Определяя тональность фразы с отрицанием, алгоритм последовательно перебирает пункты следующего списка, пока не встретит пункт, условие которого выполняется для анализируемой фразы, после чего определяет тональность в соответствии с этим пунктом:

1. Если отрицаемая часть положительна, то тональность всей фразы отрицательна; если отрицаемая часть отрицательна, то тональность всей фразы положительна.
2. Если отрицаемая часть — подлежащее или определение, то тональность всей фразы нейтральна, например:
   - Не все в правящей партии согласны с решением Асо (отрицаемая часть — подлежащее «все»).
   - «Прощание в Стамбуле» и пока не изданные «Гавани Луны» — очень североамериканские романы, очень любовно-криминальные, очень мейлеровские; «Табор уходит», напротив, очень латиноамериканский (отрицаемая часть — определение «изданные»).
   - Вивек Кундра, однако не единственный помощник Обамы по IT-вопросам (отрицаемая часть — определение «единственный»).

   Это условие можно интерпретировать так: подлежащее и определение в предложении, в первую очередь, задают тему, которой посвящено предложение, поэтому если в них и отрицается какая-либо нейтральная информация, то фраза остаётся нейтральной и не влияет на общую тональность предложения.
3. Если в отрицаемой части присутствует существительное, то тональность всей фразы отрицательна, например:
   - Делегация США была не в состоянии прокомментировать вопросы, касающиеся предполагаемых секретных тюрем (отрицаемая часть — «быть в состоянии»).

- Дело в том, что птицам стало не хватать пищи (отрицаемая часть — «стало хватать пищи»).
- В фильме нет ощущения реальности происходящего (отрицаемая часть — «ощущение реальности»).

Это условие можно интерпретировать так: сказуемые и дополнения сообщают о каких-либо свойствах происходящих процессов, действий, и если в них встречается существительное с отрицанием, то такая фраза чаще всего выражает мысль автора о том, что этот объект для действия важен, но он отсутствует, что придаёт фразе отрицательную окраску; поскольку основной смысл передаётся именно сказуемым — ремой предложения, то и сообщение об отсутствии какого-либо объекта часто приводит к формированию отрицательной тональности. Особым случаем является последний пример: в нём отрицание «нет» само по себе является сказуемым. Для публицистического стиля вообще нехарактерна такая краткость и категоричность, поэтому если в нём встречается такое построение предложения, то это является признаком того, что автор ожидал наличие какого-то объекта и для него крайне важным оказалось то, что он отсутствовал, что характеризует отрицательную тональность. Можно сказать, что такое применение отрицания — доведённый до крайности вариант с существительным в составе сказуемого, в котором отрицательное отношение выражалось через то, что для какого-либо процесса не хватало объекта, а в случае, когда отрицание само по себе является сказуемым, это показывает, что отсутствие объекта важно для нарратива как такового.

4. Если в отрицаемой части отсутствует краткое причастие, деепричастие или глагол в личной форме, то тональность всей фразы нейтральна, например:
   - Предложенное решение поддержано не всеми: некоторые астрономы предлагают провести черту за Нептуном, а ледяные карлики наподобие Плутона отнести к транснептуновым объектам в поясе Койпера (отрицаемая часть — местоимение «всеми»).
   - 31 год — возраст для самолёта не маленький (отрицаемая часть — прилагательное «маленький»).

   В то же время, наличие в отрицаемой части краткого причастия, деепричастия или глагола в личной форме может свидетельствовать об отрицательной тональности:
   - Объяснений от Павла мы до сих пор не получили (в отрицаемой части присутствует глагол в личной форме «получили»).
   - Замминистра напомнил, что доказательств наличия российских войск на Украине так и не было представлено (в отрицаемой части присутствует глагол в личной форме «было»).
   - Подготовленный план так и не был принят (в отрицаемой части присутствует краткое причастие «принят»).

   Эта часть правила позволяет отделить фразы, в которых автор сообщает, что упоминаемый объект не совершил какого-либо действия, так как такие фразы могут оказаться отрицательными, от всех остальных фраз с отрицаниями, для которых нет причин считать их выражающими отрицательное авторское отношение.
5. Если отрицаемая часть — составное сказуемое или обстоятельство, то тональность всей фразы отрицательна, например:
   - Мы должны ясно сказать руководству России, что она не может рассчитывать на партнёрство с Западом (отрицаемая часть — составное сказуемое «может рассчитывать»).
   - Во дворах всегда есть какой-то участок, который дворники не хотят убирать, потому что не могут решить, кому он принадлежит (отрицаемая часть — составное сказуемое «хотят убирать»).

- Небольшие независимые издательства, которые отчасти выполняют эту роль, живут за счёт государственных грантов, не рассчитывая на рынок (отрицаемая часть — обстоятельство «рассчитывая»).
- В 1842 году мать Натальи Дмитриевны умрёт, не встретившись с дочерью (отрицаемая часть — обстоятельство «встретившись»).

Это условие можно интерпретировать так: с помощью составного глагольного сказуемого автор обычно сообщает о характеристиках какого-либо действия, например, возможно ли оно, желаемо ли оно, поэтому отрицание при такой фразе чаще всего свидетельствует о том, что упоминаемое действие не соответствует какой-то ожидаемой или желаемой характеристике, и с помощью него автор выражает своё отрицательное отношение. С помощью составного именного сказуемого автор обычно сообщает о свойстве объекта-подлежащего; если в публицистическом тексте автор явно делает целью высказывания (ремой предложения) сообщение о свойстве объекта и категорично употребляет именно отрицание, то это чаще всего свидетельствует, как упоминалось ранее, о том, что автор ожидал, что подлежащее будет обладать каким-либо свойством, но это оказалось не так, и это часто служит признаком отрицательного авторского отношения. Аналогичная ситуация с отрицаниями в обстоятельствах — они часто выражают отрицательное авторское отношение за счёт того, что явно говорят о том, что действие-сказуемое не обладает каким-то свойством, которого можно было бы ожидать.

6. Если в отрицаемой части присутствует краткое причастие, то тональность всей фразы нейтральна, например:
   - Объяснений от Павла нами получено не было (отрицаемая часть — «было получено»).
   - Все книги из собрания Баварской библиотеки, которые более не защищены авторским правом, будут переведены в цифровой формат (отрицаемая часть — «защищены»).
   - Пик работы АСВ не пройден (отрицаемая часть — «пройден»).
7. Если отрицаемая часть — простое глагольное сказуемое (т. е. ни одно из предыдущих условий не выполняется), то тональность всей фразы отрицательна, например:
   - Объяснений от Павла мы не получили (отрицаемая часть — «получили»).
   - Я не верю никакому телевидению, я не слушаю никакого радио и не читаю никаких газет (отрицаемые части речи — «верю», «слушаю», «читаю»).
   - Руду и книги все эти имиджмейкеры и спичрайтеры точно не производят (отрицаемая часть — «производят»).

   Эта часть правила основывается на учёте разницы между активным и пассивным залогом — в случае пассивного залога, выраженного с использованием краткого причастия, в предложении публицистического стиля фраза часто служит простым сообщением о том, что оно не совершалось; предложение становится близким к констатации факта, к официально-деловому стилю, который преимущественно нейтрален. Тогда как в случае использования автором активного залога предложение становится не таким «отстранённым», чаще явно видится, что автор не равнодушен к тому, о чём говорит, что видно, если сравнить высказывания «Объяснений от Павла нами получено не было» и «Объяснений от Павла мы не получили» — во втором гораздо более явно видна авторская позиция, выражающаяся в том, что объяснений, очевидно, ждали, но Павел их не предоставил.

Точность определения правилом тональности итоговой тональности фразы составила 0.64, что значительно выше, чем у обоих более простых правил, ориентировавшихся только на тональность отрицаемой части.

Метрики качества анализа тональности предложений и матрица ошибок с новым правилом определения тональности фраз с отрицаниями приведены в таблицах 4 и 5. Поскольку с новым правилом F_1-мера выросла на 1 %, новое правило включено в алгоритм. Однако нужно отметить,

Table 4. Sentiment classification performance with negation phrases sentiment detection rule

Таблица 4. Качество работы алгоритма с внедрённым правилом выведения тональности фраз с отрицаниями

Класс предложений	Точность	Полнота	F_1-мера	Количество предложений
Положительный	0.47	0.41	0.44	533
Нейтральный	0.76	0.78	0.77	2459
Отрицательный	0.72	0.70	0.71	1495
Среднее	0.65	0.63	0.64	4487
Взвешенное среднее	0.71	0.71	0.71	4487

Доля правильных ответов алгоритма (accuracy) = 0.71

Table 5. Sentiment classification confusion matrix with negation phrases sentiment detection rule

Таблица 5. Матрица ошибок алгоритма с внедрённым правилом выведения тональности фраз с отрицаниями

предсказ. / реальн.	Положит.	Нейтр.	Отрицат.	Всего
Положительная	220	244	69	533
Нейтральная	184	1926	349	2459
Отрицательная	63	379	1053	1495

что, хотя для класса положительных предложений точность и полнота также увеличились, они всё ещё остались достаточно низкими, а значительное число предложений с отрицательной тональностью всё так же определяются алгоритмом как нейтральные.

2.3. Расширение тонального словаря

Поскольку алгоритм ошибочно определяет значительную долю тональных предложений предложений как нейтральные, для построения качественного алгоритма необходимо улучшить используемый тональный словарь. Поскольку экспертная оценка показала, что состав слов, которым в KartaSlovSent назначена отрицательная метка тональности, достаточно точен и при этом значительно шире, чем в RuSentiLex-2017, было выдвинуто предположение, что качество обнаружения отрицательных предложений может быть повышено, если дополнить используемый словарь RuSentiLex-2017 словами с отрицательной тональностью из KartaSlovSent.

Для этого из KartaSlovSent были выбраны слова, для которых доля разметчиков, которые не смогли определить тональность слова (dunno) составила не более 0.2, агрегированный показатель тональности (value, изменяется от -1 до 1) составил максимум -0.75, а показатель расхождения оценок между разметчиками (pstvNgtvDisagreementRatio) составил не более 0.05. Таких слов оказалось 5853, среди них 2414 не входили в RuSentiLex-2017; эти 2414 слов были добавлены в тональный словарь.

Метрики качества анализа тональности и матрица ошибок для алгоритма с расширенным тональным словарём приведены в таблицах 6 и 7. Поскольку полнота определения предложений с отрицательной тональностью за счёт расширения словаря возросла и вместе с тем увеличилась и средняя F_1-мера, расширенный словарь был включён в итоговую версию алгоритма.

2.4. Автоматически подобранный набор правил рекурсивного выведения тональности

Используемый в исходном алгоритме набор рекурсивных правил выведения тональности фразы по тональностям её составляющих составлен для упрощённой грамматики всего с 5 типами синтаксических единиц — например, в нём не отличаются прямые дополнения от косвенных. Была предпринята попытка повысить качество определения тональности алгоритмом за счёт построения набора правил для грамматики с 10 типами синтаксических единиц, в соответствии с которой работает построитель дерева синтаксических единиц [15]. Такой набор может позволить более качественно определять итоговую тональность за счёт того, что для разных типов синтаксических единиц будут использоваться различные правила и появится возможность учитывать больше зако-

Table 6. Sentiment classification performance with extended sentiment dictionary

Таблица 6. Качество работы алгоритма с расширенным тональным словарём

Класс предложений	Точность	Полнота	F_1-мера	Количество предложений
Положительный	0.49	0.41	0.45	533
Нейтральный	0.77	0.77	0.77	2459
Отрицательный	0.72	0.75	0.73	1495
Среднее	0.66	0.65	0.65	4487
Взвешенное среднее	0.72	0.72	0.72	4487

Доля правильных ответов алгоритма (accuracy) = 0.72

Table 7. Sentiment classification confusion matrix with extended sentiment dictionary

Таблица 7. Матрица ошибок алгоритма с расширенным тональным словарём

предсказ. / реальн.	Положит.	Нейтр.	Отрицат.	Всего
Положительная	220	244	69	533
Нейтральная	184	1926	349	2459
Отрицательная	63	379	1053	1495

номерностей русского языка, например, то, что прямое и косвенное дополнение могут по-разному влиять на тональность — тональность фраз «остановить бомбардировки» и «остановить бомбардировками», очевидно, различна.

Поскольку количество возможных в грамматике сочетаний синтаксических единиц при этом значительно больше (22 вместо 6) и составить все соответствующие правила вручную с помощью экспертов становится сложнее, набор правил подбирался автоматически. Для каждой пары синтаксических единиц необходимо было вывести 7 правил: по одному правилу для каждой из шести возможных комбинаций несовпадающих меток тональности и одно правило для случая, когда тональность обеих частей синтаксической единицы отрицательна. Такое решение обусловлено тем, что в случае, когда обе синтаксические единицы положительны или нейтральны, тональность полученной в результате их соединения фразы практически всегда является, соответственно, положительной или нейтральной, тогда как для отрицательной тональности это далеко не всегда верно, например, фраза «ненавидеть лжецов» отрицательной не является, хотя обе её части и отрицательны.

Для того, чтобы выбрать среди всех возможных наборов правил тот, который лучше всего отражает существующие в русском языке закономерности проявления тональности, использовалась кросс-валидация. Весь корпус был разбит на 5 частей, и было сформировано 5 пар обучающих и валидационных наборов предложений. Для каждого из обучающих наборов был определён набор правил, обеспечивающий максимальную F_1-меру на обучающем наборе; после этого среди пяти таких наборов был выбран тот, для которого наивысшей оказалась F_1-мера на соответствующем валидационном наборе. Существенного снижения показателей качества на валидационных наборах по сравнению с обучающими отмечено не было.

Метрики качества и матрица ошибок алгоритма с автоматически подобранным набором правил рекурсивного выведения тональности приведены в таблицах 8 и 9. Поскольку среднее качество классификации увеличилось, а существенного снижения не было замечено ни для одного из классов, исходное предположение было признано верным, и новый набор правил был включён в алгоритм.

2.5. Отдельные правила для обработки различных средств выражения тональности

Правила, описанные в данном разделе, были построены на основе знаний эксперта-филолога (одного из авторов работы) относительно средств выражения тональности, известных русской лингвистике, и выбора тех из них, которые достаточно формализованы для того, чтобы их можно было

Table 8. Sentiment classification performance with novel recursive sentiment detection ruleset

Таблица 8. Качество работы алгоритма с новым набором правил рекурсивного выведения тональности

Класс предложений	Точность	Полнота	F_1-мера	Количество предложений
Положительный	0.50	0.44	0.47	533
Нейтральный	0.79	0.78	0.78	2459
Отрицательный	0.73	0.77	0.75	1495
Среднее	0.67	0.66	0.67	4487
Взвешенное среднее	0.73	0.74	0.73	4487

Доля правильных ответов алгоритма (accuracy) = 0.74

Table 9. Sentiment classification confusion matrix with novel recursive sentiment detection ruleset

Таблица 9. Матрица ошибок алгоритма с новым набором правил рекурсивного выведения тональности

предсказ. / реальн.	Положит.	Нейтр.	Отрицат.	Всего
Положительная	235	233	65	533
Нейтральная	179	1912	368	2459
Отрицательная	55	284	1156	1495

алгоритмизировать с использованием генерируемых в рамках алгоритма деревьев синтаксических единиц.

1. Обработка синтаксических единиц (фраз) с частицей «бы».
   - Сама по себе частица «бы» не влияет на тональность фразы.
   - Синтаксическая единица, одна из частей которой «хотелось бы» или «хорошо бы», нейтральна, например (здесь и далее подчёркнута синтаксическая единица, тональность которой определяется):
     – Конечно, очень хотелось бы сказать, что открылись ворота, и к нам хлынули прекрасные фильмы Вуди Аллена или Фрэнсиса Копполы, но чёрта с два.
     – А во второй части мне хотелось бы написать про подготовку к отъезду и рассказать подробно о самой школе Marktoberdorf Summer School.
     – Хотелось бы самому посмотреть сочинские объекты.
     – Конечно, против Китая хорошо бы дружить напрямую с США, у которых с военной мощью всё хорошо.

     Во всех предложениях фраза, вводимая с помощью «хотелось бы», нейтральна: в первом предложении отрицательная тональность выражается с помощью фразеологизма «чёрта с два», второе и третье предложения нейтральны. Можно объяснить это тем, что конструкция «хотелось бы» употребляется, как правило, просто для сообщения о каком-то событии, и реальное отношение, если оно присутствует, автор выражает с помощью других фраз и конструкций. Само «хотелось бы» может как просто сообщать о намерении, например, «мне хотелось бы поехать в Дубай», так свидетельствовать о намерении, которое не осуществилось, например, «мне хотелось бы поехать в Дубай, но еду в Геленджик» — но в этом случае важно противительное придаточное с союзом «но»; с помощью «хотелось бы» автор может даже сказать о том, что его намерение осуществилась — «мне хотелось бы написать и я пишу». Используя оборот с «хорошо бы», автор выражает положительное отношение к возможному событию, но не говорит о том, реализовалось ли оно, и может ли реализоваться вообще, поэтому такую фразу в контексте анализа тональности предложений предпочтительнее считать нейтральной.
   - Синтаксическая единица, одна из частей которой — «лучше бы» или «лишь бы», имеет отрицательную тональность, например:

  - Он лучше бы за подчинёнными следил.
  - Лишь бы это не привело к печальным итогам.
  - Правительства склонны делать всё, лишь бы смотреться не хуже других.

  Фраза «лучше бы» выражает, что сейчас кто-то поступает неправильно, и автор этим эмоционально недоволен. Отличие «лучше бы» от «хорошо бы», в первую очередь, в том, что «хорошо бы» просто сообщает о некотором желаемом событии, а «лучше бы» явно сообщает, что сейчас дела идут не так, как хотелось бы автору. «Лишь бы», как и другие конструкции с «бы», придаёт неуверенность, говорит о том, что действие ещё не случилось, отчасти формирует сослагательное наклонение. «Лишь», в свою очередь, показывает отрицательное отношение к действию, что оно либо является не лучшим, с точки зрения автора, исходом, как в первом примере, либо что оно само по себе не ценно, тогда как усилия, затрачиваемые на него, чрезмерны, как во втором примере.
- Синтаксическая единица, состоящая из двух частей, в состав одной из которых входит «хотя бы», а другой — «могли бы», имеет отрицательную тональность, что обусловлено вкладываемым смыслом, заключающимся в том, что надежды автора не оправданы, например:
  - Хотя бы «Матрицу» авторы могли бы посмотреть.

2. Тональность синтаксической единицы, одна из частей которой — синтаксическая группа глагола «отобрать» или образованного от него слова, а вторая — группа косвенного дополнения с предлогом «у», отрицательна, например:
   - Спор возникает из-за отобранных у Мексики территорий.
   - Каждого такого бизнесмена я бы обложила специальным целевым пенсионным налогом, из которого делала бы доплаты к пенсии всем людям, работавшим в советский период, чтобы хотя бы частично скомпенсировать то, что у них отобрали.

   Глагол «отобрать» ближе к разговорной речи, употребляя его в публицистическом тексте, автор явно показывает, что лишение кого-либо чего-либо произошло не по доброй воле и достаточно грубо, и он не считает это лишение правильным.
3. Тональность фразы, описывающей переход из одного состояния в другое, зависит от тональностей исходного состояния и результата изменений.
   - Если результат имеет отрицательную тональность, то тональность фразы также отрицательна, например: «Поэтому вся переаттестация превратилась в фарс».
   - Если результат имеет положительную тональность, то тональность фразы также положительна, например: «Практика динамической медитации преобразует гнев в сострадание».
   - Если результат имеет нейтральную тональность, а исходное состояние — положительную, то тональность фразы отрицательна, например: «Акунин из автора стильной детективной прозы превратился в обыкновенного массового писателя».
   - Если результат имеет нейтральную тональность, а исходное состояние — отрицательную, то тональность фразы положительна, например: «Индонезия превратилась из диктатуры в нормальную страну».
   - Если тональности исходного состояния и результата совпадают, то фраза имеет такую же тональность, например: «Недавно страна превратилась из экспортёра нефти в покупателя».

   Правило применяется к синтаксической единице, одна из частей которой — синтаксическая группа сказуемого «превратить», «превратиться» или «преобразовать», а вторая — косвенное дополнение с предлогом «в». Это косвенное дополнение называет результат изменений. Если у сказуемого нет возвратного суффикса, то исходное состояние — прямое дополнение в составе синтаксической группы сказуемого («превращает диктатуру в нормальную страну»);

если же у сказуемого есть возвратный суффикс, то исходное состояние — косвенное дополнение с предлогом «из» в синтаксической группе сказуемого («превратилась из диктатуры в нормальную страну»), а если его нет, то исходное состояние — подлежащее («диктатура превращается в нормальную страну»).

4. Тональность фразы с глаголом «позволять» и образованными от него словами, сообщающей о возможности некоторого нейтрального действия, положительна, если называется тот, у кого есть эта возможность, например:
   - Сделка позволяет компании получать средства для своих сервисов.
   - Размещение научной статьи в журнале открытого доступа позволит учёному заявить об авторстве на идею.
   - Это разработка, позволяющая любому человеку написать программу для мобильной платформы Android.

   В то же время, если объект, у которого существует возможность, не называется, то такая фраза нейтральна:
   - Новые методы позволяют определять присутствие синтетического тестостерона в организме.
   - Реклама корректирующей жидкости Tipp-Ex позволяет выбирать сценарий ролика и писать его продолжение «самостоятельно».
   - Telegram — бесплатный мессенджер для смартфонов, позволяющий обмениваться текстовыми сообщениями и файлами.

   Правило применяется к синтаксической единице, одна из частей которой — группа составного глагольного сказуемого со вспомогательной частью «позволять» или группа определения «позволяющий», к которому присоединено косвенное дополнение, выраженное глаголом в начальной форме, а вторая — группа косвенного дополнения в дательном падеже без предлога.

   Это правило отражает своеобразную «солидаризацию» автора предложения с тем, у кого возникает возможность, когда он явно сообщает, у кого именно она появилась; при сообщении же о возможности, которая просто есть у неназываемого лица, такой «солидаризации» не происходит. Разница в семантике невелика, однако при нейтральной тональности она может привести к разнице в итоговой тональности фразы.
5. Тональность фразы, сообщающей о том, что оцениваемое нейтрально событие привело к положительным последствиям, положительна, например: «Сегодня наши усилия привели к успеху». Правило применяется к синтаксической единице основы предложения с нейтральным подлежащим и сказуемым «привести», в группу которого входит косвенное дополнение с предлогом «к», тональность которого положительна.
6. Тональность синтаксической единицы, одна часть которой — глагол «использовать» или «искать» или образованное от него слово, а вторая — группа подлежащего или прямого дополнения с положительной тональностью, нейтральна, например:
   - Точность GPS до миллиметров используется в гражданских целях.
   - Я ищу актрису, которая идеально подходит по образу.

   Данное правило отражает то, что глагол «использовать» часто употребляется в информационных сообщениях, когда автор пытается сообщить о факте, а не выразить своё отношение к нему; а с помощью глагола «искать» автор сообщает, что желаемое ещё не найдено, и даже если к этому желаемому он относится положительно, радость по его поводу будет преждевременна.

7. Тональность синтаксической единицы, одна часть которой — группа глагола «измерять» или «изучать» или образованного от него слова, а вторая — подлежащее либо прямое дополнение, нейтральна, например:
   - В процессе демонстрации измерялась деградация зрительной коры головного мозга.
   - Можно решать обратную задачу: напрямую измерить удовлетворённость жизнью и посмотреть, насколько она соответствует индексам.
   - Он начал изучать изобразительное искусство в 70-х годах, посещая занятия в нескольких арт-колледжах США.

   Это правило было введено, поскольку, если автор сообщает о том, что какое-то явление измерялось или изучалось, то он, скорее, рассматривает его с позиции объективного исследователя, таким образом, такое упоминание явления будет скорее нейтральным, и оно не влияет на итоговую тональность предложения.
8. Синтаксические группы слов «понятие», «концепция» или «определение» нейтральны, например:
   - Данная форма эволюции элит соответствует концепции навязывания.
   - Для Запада понятия «демократия», «свобода слова», «независимый суд», «права человека» и т. п. сверхценностны.

   Данное правило отражает то, что, если автор предложения не называет явление напрямую, а говорит, например, о его «концепции», то он показывает, что относится к нему как исследователь, беспристрастно, и такая фраза не будет выражать ни положительное отношение автора, ни отрицательное.
9. Фразы, состоящая из группы дополнения «на тему», нейтральны вне зависимости от тональности остальных составляющих группы, например:
   - Ранее на тему допинга в России высказывались прыгунья в длину Джейд Джонсон, бегун Дэй Грин и главный тренер сборной Великобритании по легкой атлетике Петер Эрикссон.
   - Вице-премьер Дмитрий Рогозин провёл совещание на тему решения проблем с выплатой заработных плат.
10. Составные сказуемые с глаголом «является» нейтральны, если их главная часть положительна или нейтральна, например: «Наиболее знаменитым продуктом компании является усилитель NAD 3020». Это правило отражает то, что глагол «является» свойственен скорее научному или официально-деловому стилю речи и, используя его, автор скорее отстраняется от того, чем именно является упоминаемый объект.
11. Фраза-сообщение о том, что объект обладает каким-либо нейтральным свойством, обладает положительной тональностью, если она является предикативным ядром (основой) предложения, например:
   - Сейчас правящая коалиция обладает большинством в Кнессете.
   - Усилители NAD обладают звуком, который обычно присущ аппаратам более высокой ценовой категории.

   В то же время, если сообщение об обладании не является предикативным ядром предложения, то оно нейтрально:
   - Израиль — ближневосточная страна, обладающая ядерным оружием.

   Правило отражает то, что само по себе слово «обладать» является достаточно возвышенным, относящимся скорее к литературно-художественному стилю, и, если автор делает на нём акцент, вынося его в основу предложения, то это признак того, что, с точки зрения автора, факт обладания важен, и, если нет оснований считать такую фразу отрицательной, то она скорее выражает положительное авторское отношение.

12. Синтаксическая единица, одна из частей которой — прилагательное «необходимый» или образованное от него слово, нейтральна, например:
    - Для завершения программ QE необходим сильный рынок труда, пишет РБК.
    - Для использования сервиса будет необходимо заменить действующую SIM-карту телефона на специальную SIM-карту с поддержкой технологии NFC.

    Это правило отражает, что оценка «необходим» используется, как правило, в формальных или информационных сообщениях, и автор, используя его вместо разговорного «нужен», как бы отстраняется от сказанного и показывает, что эта оценка объективна, и он не имеет какого-то своего мнения по теме.
13. Тональность синтаксической единицы основы предложения, подлежащее которой — глагол в начальной форме, а сказуемое — глагол «приходится», отрицательна, например:
    - Требования в ВУЗах мягкие и боевой опыт приходится приобретать в боевой же обстановке.
    - Действовать в правовом поле приходится методом проб и ошибок.
    - Я хочу ходить в клубы потанцевать, но в итоге приходится танцевать только или дома для себя, или в узкой компании друзей.

    В таких предложениях автор явно указывает на то, что совершение действия против воли, что выражает отрицательное авторское отношение.
14. Синтаксическая единица основы предложения, подлежащее которой — глагол в начальной форме, а сказуемое — предикативом «нельзя», нейтральна, например:
    - Пока нельзя предполагать, что США отключат GPS.
    - Нельзя исключать появление мощного исламистского фронта непосредственно у границ России.
    - Вторую накопительную башню, находящуюся на улице Советской, назвать нельзя действующей.

    В публицистическом стиле такие предложения чаще всего нейтральны, они сообщают, что автор не имеет чёткого мнения по вопросу, и не хочет настаивать на какой-либо точке зрения.
15. Тональность фраз, сообщающих о противостоянии, борьбе с чем-либо, положительна, если борьба происходит с явлением, которое описывается отрицательно, отрицательная, если с явлением, описывающимся положительно, иначе нейтральной, например:
    - Тренировка убирает навязчивые черты.
    - Сейчас уменьшается количество источников правдивой информации.
    - Это будет препятствовать производству героина.
    - Мы разрушили планы террористов по проведению новых атак.
    - Традиционно любимые согражданами книги в переплёте вытесняются более дешёвыми книжками в мягкой обложке.
    - Защитников флоры атаковали около 05:00 примерно полсотни неизвестных в масках, похожих на ультраправых футбольных фанатов.
    - Потому что стальной брони нет, а эти удары ослабляют защиту от самых страшных болезней.
    - Поправки должны избавить рынок от нелегальных букмекеров и защитить игроков от мошенников.
    - И Япония, и Россия хотят, чтобы эта проблема была решена.

    Правило применяется к синтаксической единице, удовлетворяющей одному из условий:
    - одна из частей — синтаксическая группа одного из глаголов «атаковать», «вытеснять», «ослаблять», «препятствовать», «разрушать», «решать», «убирать», «угрожать», «уменьшать», «не допускать» или образованного от него слова, а вторая — прямое дополнение,

- одна из частей — синтаксическая группа одного из глаголов «защищать», «избавлять», «расчищать», а вторая — косвенное дополнение с предлогом «от».
- одна из частей — синтаксическая группа глагола «бороться» или образованного от него слова, а вторая — косвенное дополнение с предлогом «с»,

16. Фразы, сообщающие о том, что какому-либо действию предшествовало отрицательное событие, нейтральны и не учитываются при дальнейшем выведении тональности, например:
    - За 17 лет, прошедших после трагедии распада СССР, обе страны много выиграли от взаимовыгодного сотрудничества.
    - Главный герой — консьерж месье Густав (Рэйф Файнс), который после смерти одной из постоялиц получает в наследство бесценную картину.

    Правило применяется к синтаксической единице, одна из частей которой — синтаксическая группа сказуемого или определения, выраженного причастием, а вторая — синтаксическая группа обстоятельства с предлогом «после».
17. Фразы, содержащие авторскую положительную или отрицательную оценку успешности действия, являются, соответственно, положительными или отрицательными вне зависимости от самого действия, например:
    - Бортовой лазер успешно уничтожил баллистическую ракету.
    - Неудачно выступил на главном турнире года Владимир Крамник.

    Данное правило было отражает, что само по себе описание действия может содержать как положительную, так и отрицательную лексику («уничтожил», «главный турнир»), но, если сам автор явно выносит действию ту или иную оценку, то она гораздо лучше отражает его авторскую позицию, и, следовательно, именно она должна использоваться для определения тональности всей фразы.
18. Фразы, содержащие простое сообщение о мнении третьего лица, без оценки этого мнения автором предложения, нейтральны, например:
    - По мнению аналитиков, сделка перспективна для обеих компаний.
    - По мнению многих немцев, сборник Liebe ist für alle da пропагандирует порнографию, насилие и незащищённый секс.
    - Мнение Юрия Любимова о том, что нужно делать, чтобы театр не превратился в музей самого себя, что должен уметь режиссёр и как правильно разговаривать с актёрами, читайте далее.
    - Министр отметил, что борьба с допингом сейчас крайне важна.
    - Теория гласит, что получившиеся гибридные кролики были особенно выносливыми и энергичными.
    - Интернет-издание Engadget сообщает, что энтузиасты уже опробовали новую операционную систему Google Chrome OS, сделав загрузочный диск на обычной флешке.
    - Н. Кан заверил, что перестановки в правительстве помогут укрепить лидерство партии и создать сильную команду, способную реформировать страну.

    Правило отражает то, что в классическом публицистическом тексте автор предложения с помощью конструкции «по мнению», «по словам», «считает», «сообщает», «заверяет», «гласит» и аналогичных явно указывает на то, что он лишь сообщает о том, что думает какой-то человек, не выражая своего согласия или несогласия с оценками этого человека. Кроме того, такая конструкция сама по себе подразумевает, что автор не может утверждать, что дела действительно обстоят так, как кто-то говорит, наоборот, это в первую очередь причина продолжить разбор темы и поговорить о ней ещё, прежде чем вынести свою оценку. Также использование подобных фраз может сообщать, что автор предложения не уверен, что дела обстоят именно так, как говорится, и сообщает не о самом факте, а то, что у кого-то есть мнение: например,

в предложении «По мнению руководства Северной Кореи, Сеул и Вашингтон сфабриковали обвинения против Пхеньяна» автор сообщает не о сфабрикованности обвинений, иначе тональность была бы отрицательной, а о том, что руководство КНДР считает обвинения сфабрикованными; автор не говорит, так это или нет на самом деле.

Однако сообщение о мнении самого автора может быть тональным, например:

- Отметим, что резкое обвинение Юлхяэ может стать поводом для того, что российским спортсменам в Ванкувере попросту не дадут спокойно выступать, замучив их бесконечными проверками на пробы.
- Важно отметить, что наше сотрудничество с Пакистаном помогло нам выследить бен Ладена и отследить здание, в котором он скрывался.
- Следует отметить, что на слухах и сплетнях вокруг предстоящего через три года конца света нажились тысячи мошенников по всему миру.
- Я знаю, что если выделить в одной большой задаче 10 маленьких подзадач, то после этого работа идёт как по маслу.

Алгоритм считает синтаксическую единицу изъяснительного придаточного нейтральной, если для неё выполняется одно из условий:

- она присоединяется к одному из глаголов «гласить», «заверять», «знать», «отмечать», «сообщать» или «считать», стоящем в форме третьего лица,
- она присоединяется к слову «мнение», в синтаксическую группу которого не входят притяжательные местоимения первого лица («моё», «наше»).

19. Предложения, построенные по газетному шаблону «подробнее о… читайте в… », всегда нейтральны, например:
    - Подробнее о развитии конфликта на Корейском полуострове читайте в статье Частного корреспондента «Затишье перед бурей».
    - Подробнее о претензиях к телеканалу «Дождь» читайте в статье «Не столько дождь лил, сколько гром гремел».
20. Тональность фраз, содержащих изъяснительное придаточное, придаточное причины или следствия, либо определительное придаточное, присоединённое с помощью конструкции «в том, что», определяется тональностью придаточного, вне зависимости от правила рекурсивного выведения.

    Изъяснительные придаточные:
    - Уже первые минуты игры показали, что эти опасения были напрасны.
    - Это новости, которые доказывают, что даже в бедных странах существуют условия для свободного выражения мнений.
    - Два месяца опыта показывают, что лучше этой платформы на данный момент нет.

    Придаточные причины:
    - Александр сказал, что приедет ещё, потому что ему очень понравилось.
    - Исторических потому, что мы всему миру ещё раз доказали демократичность нашего общества.

    Придаточные следствия:
    - А вот «Дозоры» Лукьяненко я не читал, поэтому первый фильм посмотрел без содроганий.
    - У нас в роду сплошь мальчики, поэтому Дашенька — это большое счастье и подарок, особенно для нас с дедушкой.
    - Число генералов сократилось, поэтому можно говорить о положительных тенденциях, привнесённых переаттестацией.

- Нет закона о социальном предпринимательстве, поэтому действовать в правовом поле приходится методом проб и ошибок.

Определительные придаточные, присоединённые с помощью конструкции «в том, что»:

- Дело в том, что в программе «Кинотавра» было много хорошего кино.
- Суть сервиса в том, что он позволит создавать программы людям, которые в программировании ничего не понимают.
- Особенность усилителей NAD состоит в том, что они обладают качеством звука, который обычно присущ аппаратам более высокой ценовой категории.

21. Синтаксическая единица вопросительного предложения с положительной основой нейтральна, например:
    - Довлатов — это классик или современник?
    - Как удаётся отвлечься, чтобы удачно нырнуть в океан и поймать рыбу?
    - На чём же вы выиграть хотите?

    Это правило отображает, что вопросительное предложение не может иметь положительную тональность, поскольку оно всегда содержит в себе долю сомнения, неуверенности в сказанном, и автор, даже используя в нём положительные оценки, всё-таки показывает, что они для него находятся под сомнением. Тем не менее, для отрицательных оценок это не так, и вопросительное предложение может иметь отрицательную тональность:
    - И почему мне не хватает терпения?
    - Как только можно сравнивать сталинизм с нацизмом?
22. Синтаксическая единица восклицательного предложения с нейтральной основой положительна, например:
    - Такого старта у нас ещё не было!
    - Оказалось, что один в поле воин!
    - Назовите кого-нибудь, кто обладает таким успехом!
    - Этим космонавтом был гражданин Советского Союза, майор Юрий Алексеевич Гагарин!

    Это правило отображает, что восклицательное предложение не может быть нейтральным, поскольку восклицательные предложения в языке служат для выражения авторской экспрессии, и, если автор его использовал, то у него точно есть своё мнение по теме, которое он хочет выразить. При этом, если основу предложения нет оснований считать положительной или отрицательной, то это, скорее всего, значит, что само событие, описываемое в основе, является для автора настолько важным, что он сообщает о нём в восклицательном предложении, что свидетельствует о положительном отношении автора.
23. Тональность синтаксической единицы, одна из частей которой — шаблонное вводное слово или вводная конструкция, выражающая отрицательное авторское отношение, такое как «увы» или «к сожалению», отрицательна вне зависимости от других встреченных средств проявления тональности, например:
    - К сожалению, местная избирательная комиссия утвердила заявление.
    - Увы, добившись успеха, группа чаще стала выступать на огромных стадионах.

    Данное правило основано на том, что вводные слова, как правило, служат для связки автором предложений и выражаемых в них мыслей в отдельный текст, поэтому, если автор явно использует вводное слово, показывающее, что он недоволен сообщаемой информацией, то именно это и выражает авторское отношение, даже если сообщаемую информацию можно было бы воспринимать положительно — например, саму по себе информацию о том, что музыкальная группа стала популярной и теперь выступает на стадионах, вполне можно рассматривать как положительную, но автор явно говорит, что он предпочёл бы, чтобы этого не происходило, и лично он относится к этому отрицательно.

Table 10. Sentiment classification performance with special syntactic rules for sentiment expression means processing

Таблица 10. Качество работы алгоритма со специальными правилами для обработки различных средств выражения тональности

Класс предложений	Точность	Полнота	F_1-мера	Количество предложений
Положительный	0.66	0.64	0.65	533
Нейтральный	0.84	0.81	0.83	2459
Отрицательный	0.77	0.82	0.80	1495
Среднее	0.76	0.76	0.76	4487
Взвешенное среднее	0.80	0.80	0.80	4487

Доля правильных ответов алгоритма (accuracy) = 0.80

Table 11. Sentiment classification confusion matrix with special syntactic rules for sentiment expression means processing

Таблица 11. Матрица ошибок алгоритма со специальными правилами для обработки различных средств выражения тональности

предсказ. / реальн.	Положит.	Нейтр.	Отрицат.	Всего
Положительная	342	156	35	533
Нейтральная	135	2000	324	2459
Отрицательная	38	225	1232	1495

Метрики качества и матрица ошибок алгоритма с правилами для обработки различных средств выражения тональности приведены в таблицах 10 и 11. Использование специальных правил привело к достаточно существенному росту качества определения тональности — средняя F_1-мера увеличилась на 0.09, до 0.76, а средняя взвешенная — на 0.07, до 0.80. Этот рост обусловлен в первую очередь значительно возросшими точностью и полнотой определения предложений с положительной тональностью — на 0.16 и 0.20 соответственно. За счёт введения правил, обрабатывающих различные средства выражения положительной тональности, алгоритм стал гораздо лучше отделять положительный класс предложений от нейтрального. Точность и полнота определения нейтральных предложений и предложений с отрицательной тональностью также увеличились как минимум на 3 %. Таким образом, добавление специальных правил обработки различных средств выражения тональности позволило достаточно сильно увеличить качество работы алгоритма.

3. Анализ ошибок и обсуждение результатов

Для определения наиболее перспективных путей дальнейшего развития алгоритма была собрана информация о 180 предложениях, тональность которых была определена неверно (по 30 предложений для каждого из 6 возможных сочетаний реальной и определённой алгоритмом тональности). Они были разделены на 11 групп в зависимости от причины неверной классификации (см. таблицу 12).

В группу с некорректной разметкой тональности попали предложения, которые, несмотря на все предосторожности при построении OpenSentimentCorpus, получили неверную разметку и при этом не попали в выборку для перепроверки. Часть из этих предложений сложны для понимания и должны были быть отмечены как сомнительные при разметке, например «Вниманием, прогулками по еловому ботаническому саду», однако некоторые просто были неверно поняты разметчиками, например предложение «По тому, что у меня в списке целых три Трифонова, вы уже поняли, кого я считаю главным советским писателем» было отмечено разметчиками как нейтральное, хотя оно имеет положительную тональность. Большинство из них были отмечены как нейтральные, что говорит о необходимости при создании корпусов в будущем уделить особое внимание разметке предложений этого класса.

В группу с некорректным построением дерева синтаксических единиц попали предложения, для которых либо некорректно отработал парсер дерева синтаксических связей, либо оказался некор-

Table 12. Distribution of errors of the proposed algorhtm in %

Таблица 12. Распределение ошибок предложенного алгоритма по группам в %

Реальная тональность	Положит.		Нейтральн.		Отрицат.		В среднем
Предсказанная тональность	Нейтр.	Отр.	Пол.	Отр.	Пол.	Нейтр.	
Некорректная разметка тональности предложения	0	10	33	20	6	6	13
Некорректное построение дерева синтаксических единиц	10	23	4	4	13	20	12
Некорректное определение тональности одиночного слова	17	4	13	7	13	20	12
Некорректное определение тональности устойчивого сочетания слов	3	10	4	3	23	17	10
Несовершенство существующих правил определения тональности	23	0	0	10	4	0	6
Отсутствие правила для обработки сообщения о чужом мнении	7	0	13	20	0	0	7
Отсутствие правила для обработки семантики противостояния, борьбы	3	23	0	3	10	3	7
Отсутствие правила для обработки семантики увеличения или уменьшения	3	0	0	3	4	10	3
Отсутствие правила для обработки придаточного предложения	17	7	3	0	7	3	6
Отсутствие правила для обработки иного средства выражения тональности	17	20	7	23	13	17	16
Тональность определяется высокоуровневой структурой предложения	0	3	23	7	7	3	7

ректен алгоритм построения деревьев синтаксических единиц. Общее количество таких предложений оказалось достаточно велико, но это, как показано в [15], неизбежно при использовании современных средств построения синтаксических деревьев.

В группы с некорректным определением тональности одиночного слова и тональности устойчивого сочетания слов попали:

- тональные предложения, тональность в которых выражается с помощью использования оценочной или эмотивной лексики, которая должна была присутствовать в тональном словаре, но из-за несовершенства словаря в него не попала, например, предложение с отрицательной тональностью «Молодая мать осталась одна», тональность в котором выражается с помощью устойчивого словосочетания «остаться одному», которое отсутствует в тональном словаре;
- нейтральные предложения с лексикой, которая была отмечена в тональном словаре как имеющая положительную или отрицательную тональность, например, нейтральное предложение «Наш герой сдёргивает с него простыню, а у того вместо тела — тело свиньи», в котором употребляется слово «герой», отмеченное в тональном словаре как имеющее положительную тональность, так как оно может использоваться для вынесения автором оценки, например, «Он — наш герой», но в анализируемом приложении нейтрально.

Эти две группы ошибок оказали достаточно существенное влияние на качество работы алгоритма, в сумме приведя к 22 % допущенных им ошибок. Нужно отметить значительное число отсутствующих в тональном словаре слов с положительной тональностью (17 % положительных предложений, ошибочно определённых как нейтральные), а также большое число сочетаний слов

с отрицательной тональностью, которые не вошли в тональный словарь (в среднем 20 % ошибок при определении отрицательной тональности).

Достаточно небольшое число предложений, ошибочно классифицированных алгоритмом из-за несовершенства существующих правил определения тональности, показывает, что введённые правила всё-таки оказались достаточно точными. Нужно отметить только большое количество (23 %) положительных предложений, ошибочно определённых как нейтральные. Вероятно, для отделения этих классов предложенные правила оказались слишком грубыми.

Правильному определению тональности многих предложений также помешало отсутствие правил для обработки некоторых средств выражения тональности, например: «Для автора красота есть признак здорового тела и гармонии со Вселенной» — автор сообщает о чужом мнении, но не показывает, согласен ли он с ним; «Динамическая медитация особенно эффективна для людей, страдающих от бессонницы» — автор сообщает о семантике борьбы медитации с бессонницей; «Отметим понижение организованности процесса голосования» — отрицательная тональность выражается с помощью оценки организованности как ухудшающейся; «Мы продолжим нашу трудную работу, чтобы сделать страну сильной и единой» — положительная тональность выражается с помощью придаточного цели. Главной проблемой создания правил для обработки таких средств оказывается их относительно редкое использование в языке, что затрудняет экспертам-лингвистам определение влияния на тональность того или иного средства выразительности. Тем не менее, эта часть ошибок представляет собой одно из важнейших направлений развития — суммарно на неё приходится 29 % случаев неправильного определения тональности. Особенно важна обработка сообщения для чужом мнении для более точного определения нейтральных предложений — из-за неё допущено в среднем 17 % ошибок для нейтральных предложений.

В группу предложений, тональность которых определяется высокоуровневой структурой, попали предложения, тональность которых не может быть определена на основе тональностей их составных частей. Например, нейтральное предложение «Людей не принуждают принимать радикальную идеологию, они сами приходят к ней в результате перемен, происходящих в обществе» состоит из двух частей: в первой сообщается, что людей не принуждают принимать радикальную идеологию, и эта часть само по себе имеет положительную тональность; во второй сообщается, что люди сами приходят к ней в результате происходящих перемен, и эта часть, если рассматривать её изолированно, также скорее положительна, однако если рассматривать всё предложение в целом, то оно нейтрально. Такие ошибки относятся к ограничениям алгоритма, и, несмотря на то, что их было допущено всего 7 % от общего количества, для нейтральных предложений, ошибочно определённых как положительные, эта проблема — одна из основных.

Подводя итоги анализа ошибок, можно сделать следующие выводы.

1. Некорректная разметка оказывает значительное влияние на качество определения тональности нейтральных предложений, для остальных двух классов она малозначима. Вероятнее всего, это связано с несовершенством руководств по разметке для нейтрального класса.
2. Некорректное построение дерева синтаксических единиц, наоборот, сильнее всего влияет на анализ тональных предложений, в особенности на отделение отрицательных предложений от двух других классов. Можно предположить, что причина этого в большей синтаксической сложности отрицательных предложений: в них чаще встречаются отрицания и противительные союзы, из-за чего эти предложения сложнее для автоматического синтаксического разбора.
3. Некорректное определение тональности одиночных слов сильнее всего влияет, во-первых, на обнаружение отрицательной тональности, во-вторых, на отделение положительных предложений от нейтральных. Некорректное определение тональности устойчивых сочетаний

слов влияет преимущественно на обнаружение отрицательной тональности. Вероятно, эта часть тонального словаря нуждается в серьёзном расширении.

4. Несовершенство существующих правил определения тональности серьёзно проявляется только в том, что многие положительные предложения определяются как нейтральные. Скорее всего, это вызвано тем, что при автоматическом подборе набора правил рекурсивного выведения положительных предложений оказалось слишком мало, чтобы сформировать для них надёжные правила.
5. Недостаточность правил для обработки сообщения о чужом мнении влияет преимущественно на класс нейтральных предложений, в результате многие нейтральные предложения ошибочно определяются как тональные.
6. Недостаточность правил для обработки семантики противостояния, борьбы влияет в первую очередь на то, что тональность многих положительных предложений определяется алгоритмом как отрицательная, а во вторую — на аналогичную ошибку для отрицательных предложений.
7. Отсутствие правил для обработки семантики увеличения или уменьшения практически не оказывает влияния на качество работы алгоритма.
8. Отсутствие правил для обработки придаточных предложений влияет, в первую очередь, на обнаружение средств выражения положительной тональности.

В целом наиболее серьёзные проблемы алгоритма — недостаточность набора слов с положительной тональностью и устойчивых сочетаний слов с отрицательной тональностью в тональном словаре, а также нехватка правил для обработки различных средств выражения положительной тональности.

Заключение

В статье рассмотрена разработка алгоритма определения тональности русскоязычных предложений, основанного на применении семантических правил и ориентированного на применение к текстами в публицистическом стиле. Алгоритм рекурсивно применяет подходящие правила к составным частям предложения, представленным в виде дерева синтаксических единиц. Большинство правил было построено на основе знаний эксперта-филолога относительно средств выражения тональности, известных русской лингвистике, и выбора тех из них, которые достаточно формализованы для того, чтобы их можно было алгоритмизировать с использованием генерируемых в рамках алгоритма деревьев синтаксических единиц. Также применялись такие инструменты, как дерево решений и тональный словарь.

В экспериментах с разработанным алгоритмом удалось достичь F_1-меры, равной 0.80, что является существенным шагом вперёд в анализе тональности предложений публицистического стиля. В более ранних работах, посвящённых данному вопросу (например, [17], в которой использовался SentiStrength) была получена F_1-мера, равная 0.60. Для английского языка существуют подходы, позволяющие добиться сравнимой с предложенным алгоритмом F_1-меры, в частности, 0.76 для новостных текстов [18] при использовании LSTM и 0.71 для предложений из LiveJournal [19] при использовании набора правил, построенного с помощью генетического программирования.

Проведённый анализ ошибок позволил идентифицировать основные проблемы алгоритма и на их основе сформулировать направления для его развития — совершенствование тонального словаря и введение новых семантических правил для обработки различных средств выражения положительной тональности.

References

[1] B. Liu, *Sentiment Analysis and Opinion Mining*. Springer, 2022, 167 pp.

[2] A. Dvoybikova, A. Karpov, and O. Verkholyak, "Analytical review of methods for identifying emotions in text data", in *3rd International Conference on R. Piotrowski's Readings in Language Engineering and Applied Linguistics, PRLEAL 2019*, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, 2020, pp. 8–21.

[3] 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.

[4] K. Nursakitov, A. Bekishev, S. Kumargazhanova, and A. Urkumbaeva, "Review of methods for determining the tonation of texts in natural languages", *Bulletin of Shakarim University. Technical Sciences*, no. 1 (9), pp. 59–67, 2023.

[5] M. S. Başarslan and F. Kayaalp, "Sentiment analysis on social media reviews datasets with deep learning approach", *Sakarya University Journal of Computer and Information Sciences*, vol. 4, no. 1, pp. 35–49, 2021.

[6] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges", *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[7] E. N. Tulupova and E. V. Golovina, "Lexico-stylistic percularities of tourist's Internet commentary", *Philology. Theory & Practice*, vol. 12, no. 5, pp. 257–261, 2019, in Russian.

[8] E. I. Boychuk, "Lexical and grammatical features of Internet reviews in the Russian and English languages", *Verhnevolzhski Philological Bulletin*, no. 3 (26), pp. 107–115, 2021, in Russian.

[9] A. Y. Poletaev and I. V. Paramonov, "Recursive sentiment detection algorithm for Russian sentences", *Automatic Control and Computer Sciences*, vol. 57, no. 7, pp. 740–749, 2023.

[10] M. Eremina, "Rechevoj zhanr otzyva v kommunikativnom prostranstve interneta", *Nauchnyj dialog*, no. 5 (53), pp. 34–45, 2016, in Russian.

[11] A. R. Kalashnikova, "Informativnaya tekstovaya tonal'nost' kak opredelyayushchij faktor ritmicheskoj tekstovoj organizacii", *Izvestiya Volgogradskogo Gosudarstvennogo Pedagogicheskogo Universiteta*, vol. 3 (107), pp. 113–116, 2016, in Russian.

[12] I. V. Paramonov and A. Y. Poletaev, "Annotation of text corpora by sentiment and presence of irony within a project of citizen science", *Modelirovanie i Analiz Informatsionnykh Sistem*, vol. 30, no. 1, pp. 86–100, 2023, in Russian.

[13] N. Loukachevitch and A. Levchik, "Creating a general Russian sentiment lexicon", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 1171–1176.

[14] D. Kulagin, "Publicly available sentiment dictionary for the Russian language KartaSlovSent", in *Computational Linguistics and Intellectual Technologies: Proceesings of the Annual "Dialog" Conference (2021)*, in Russian, 2021, pp. 1106–1119.

[15] A. Y. Poletaev, I. V. Paramonov, and E. I. Boychuk, "Algorithm of constituency tree from depencency tree construction for a Russian-language sentence", *Informatics and Automation*, vol. 22, no. 6, pp. 1323–1353, 2023, in Russian.

[16] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Routledge, 2017, 368 pp.

[17] O. Koltsova, S. Alexeeva, S. Pashakhin, and S. Koltsov, "PolSentiLex: Sentiment detection in socio-political discussions on Russian social media", in *Conference on Artificial Intelligence and Natural Language*, 2020, pp. 1–16.

[18] W. Souma, I. Vodenska, and H. Aoyama, "Enhanced news sentiment analysis using deep learning methods", *Journal of Computational Social Science*, vol. 2, no. 1, pp. 33–46, 2019.

[19] A. B. Junior, N. F. F. da Silva, T. C. Rosa, and C. G. Junior, "Sentiment analysis with genetic programming", *Information Sciences*, vol. 562, pp. 116–135, 2021.