MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 29, NO. 2, 2022
journal homepage: www.mais-journal.ru

THEORY OF DATA

Recursive Sentiment Detection Algorithm for Russian Sentences

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

DOI: 10.18255/1818-1015-2022-2-134-147

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

MSC2020: 68T50
Research article
Full text in Russian

Received April 30, 2022
After revision May 22, 2022
Accepted May 25, 2022

The article is devoted to the task of sentiment detection of Russian sentences. The sentiment is conceived as the author's attitude to the topic of a sentence. This assay considers positive, neutral, and negative sentiment classes, i.e., the task of three-classes classification is solved.

The article introduces a rule-based sentiment detection algorithm for Russian sentences. The algorithm is based on the assumption that the sentiment of a phrase can be determined by the sentiments of its parts by the recursive application of appropriate semantic rules to the sentiments of its parts organized as a constituency parse tree. The utilized set of semantic rules was constructed based on a discussion with experts in linguistics. The experiments showed that the proposed recursive algorithm performs slightly worse on the hotel reviews corpus than the adapted rule-based approach: weighted F_1-measures are 0.75 and 0.78, respectively. To measure the algorithm efficiency on complex sentences, we created OpenSentimentCorpus based on OpenCorpora, an open corpus of sentences extracted from Russian news and periodicals. On OpenSentimentCorpus the recursive algorithm performs better than the adapted approach does: F_1-measures are 0.70 and 0.63, respectively. This indicates that the proposed algorithm has an advantage in case of more complex sentences with more subtle ways of expressing the sentiment.

Keywords: sentiment analysis; sentiment detection; semantic rules; sentiment corpus

INFORMATION ABOUT THE AUTHORS

Anatoliy Yurievich Poletaev	orcid.org/0000-0003-0116-4739. E-mail: anatoliy-poletaev@mail.ru post-graduate student.
Ilya Vyacheslavovich Paramonov correspondence author	orcid.org/0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org PhD, associate professor.

Funding: The reported study was funded by YarSU Program according to the research project No. P2-GM5-2021.

For citation: A. Y. Poletaev and I. V. Paramonov, "Recursive Sentiment Detection Algorithm for Russian Sentences", *Modeling and analysis of information systems*, vol. 29, no. 2, pp. 134-147, 2022.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 29, № 2, 2022
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Рекурсивный алгоритм определения тональности предложений на русском языке

А. Ю. Полетаев[1], И. В. Парамонов[1]

DOI: 10.18255/1818-1015-2022-2-134-147

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

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 30 апреля 2022 г.
После доработки 22 мая 2022 г.
Принята к публикации 25 мая 2022 г.

В статье рассматривается задача определения тональности русскоязычных предложений. Тональность понимается как отношение автора к теме предложения. В данном исследовании учитываются три варианта тональности — положительная, отрицательная и нейтральная, т. е. решается задача классификации с тремя классами.
В статье предлагается алгоритм определения тональности предложения на русском языке, основанный на семантических правилах. В основе алгоритма лежит предположение о том, что тональность фразы может быть определена на основе тональностей её составляющих с помощью рекурсивного применения семантических правил к составным частям фразы, представленным в виде синтаксического дерева. Набор семантических правил, используемых алгоритмом, был составлен в результате обсуждений с экспертами-филологами. Эксперименты показали, что предложенный рекурсивный алгоритм даёт несколько худший результат на корпусе отзывов на отели по сравнению с подходом, основанным на правилах, ранее адаптированным авторами для русского языка: взвешенная F_1-мера составила 0.75 и 0.78 соответственно. Для оценки качества работы алгоритма на сложных предложениях был создан корпус OpenSentimentCorpus, основанный на OpenCorpora — открытом корпусе предложений из новостных статей и публицистики. На корпусе OpenSentimentCorpus рекурсивный алгоритм работает лучше, чем адаптированный подход: F_1-мера составила 0.70 и 0.63 соответственно. Таким образом, предложенный в данной работе алгоритм имеет преимущество в случае более сложных предложений с более тонкими способами выражения тональности.

Ключевые слова: анализ тональности; определение тональности; семантические правила; тональный корпус

ИНФОРМАЦИЯ ОБ АВТОРАХ

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

Илья Вячеславович Парамонов
автор для корреспонденции | orcid.org/0000-0003-3984-8423. E-mail: ilya.paramonov@fruct.org
канд. физ.-мат. наук, доцент.

Финансирование: Исследование выполнено в рамках Программы развития ЯрГУ, проект № П2-ГМ5-2021.

Для цитирования: A. Y. Poletaev and I. V. Paramonov, “Recursive Sentiment Detection Algorithm for Russian Sentences”, *Modeling and analysis of information systems*, vol. 29, no. 2, pp. 134-147, 2022.

Введение

В данной статье предлагается алгоритм определения тональности предложений на русском языке, развивающий предыдущие исследования авторов [1], и производится его оценка. Под задачей определения тональности на уровне предложений понимается выявление и классификация отношения автора предложения к его теме. Предложение считается имеющим положительную тональность, если оно содержит положительные факты, мнения или эмоции, выраженные автором, и не содержит отрицательных фактов, мнений или эмоций, либо эти отрицательные коннотации признаются автором несущественными. В противоположном случае (когда отрицательные факты, мнения или эмоции превалируют) предложение считается имеющим отрицательную тональность. Если предложение не имеет ни положительной, ни отрицательной тональности, оно считается нейтральным (имеющим нейтральную тональность) [2, 3].

Предлагаемый в настоящей работе алгоритм основан на семантических правилах. Такие подходы менее распространены по сравнению с подходами, основанными на нейронных сетях, однако в некоторых случаях они восполняют недостатки нейронных сетей: при использовании семантических правил не требуются большие обучающие корпуса, а результаты их применения достаточно легко интерпретировать. В предыдущей работе авторов [1] был адаптирован к русскому языку подход, основанный на правилах и изначально предложенный Y. Xie для английского языка [4]. Семантические правила в исходном подходе реализованы как шаблоны, применяемые к списку слов предложения. Такой метод трудно применять для русского языка, поскольку в нём, в отличие от английского, нет строгого порядка слов. Для решения этой проблемы авторы реализовали правила как алгоритмы над деревом синтаксических связей (dependency parse tree), отражающим зависимости между словами предложения. Качество работы адаптированного подхода было оценено на корпусе отзывов на отели; F_1-мера составила 0.73. Наиболее распространённый в настоящее время подход, использующий нейронную сеть BERT, даёт на этом же корпусе несколько лучший (на 5 %) результат.

Цель настоящей работы — предложить новый алгоритм определения тональности, использующий синтаксическую структуру предложения более полно, чем ранее адаптированный подход, и оценить качество его работы. Как источник информации о структуре предложения используется дерево составляющих (constituency parse tree). Алгоритм основан на предположении, что тональность фразы можно определить на основе тональностей её частей, представленных как дочерние узлы дерева составляющих. Как следствие, был разработан рекурсивный алгоритм определения тональности над деревом составляющих.

Кроме того, предполагается, что тональность предложения существенно зависит от тональности его предикативного ядра, которое является корнем дерева составляющих. Таким образом, предложенный алгоритм начинает работу от корня дерева составляющих данного предложения и рекурсивно определяет тональность каждой фразы (узла дерева составляющих), применяя семантические правила к тональностям её составных частей (дочерних узлов).

Большинство исследований, посвящённых анализу тональности русскоязычных предложений, использует корпуса, состоящие из достаточно коротких и простых предложений (например, твиты и отзывы) [5]. В таких предложениях тональность проявляется чётко и определённо. Предложения с более сложной структурой обычно не рассматриваются. Чтобы закрыть этот пробел, авторами работы был создан корпус OpenSentimentCorpus, состоящий из предложений, извлечённых из корпуса новостных и публицистических текстов OpenCorpora. Эффективность работы алгоритма была оценена и сопоставлена с эффективностью ранее адаптированного авторами подхода как на корпусе отзывов на отели, так и на OpenSentimentCorpus.

Оставшаяся часть работы организована следующим образом. В разделе 1 проведён обзор связанных работ. В разделе 2 приведено общее описание предлагаемого рекурсивного алгоритма. Раздел 3

посвящён новому корпусу, использованному для экспериментов. Раздел 4 содержит результаты экспериментов: оценку качества работы предложенного алгоритма и её сравнение с результатами ранее адаптированного подхода. В разделе 5 проведён анализ ошибок рекурсивного алгоритма на корпусе OpenSentimentCorpus. В заключении подведены итоги работы и намечены направления дальнейших исследований.

1. Обзор связанных работ

Использование семантических правил для анализа тональности на уровне предложений было предложено в работе [6] и позднее усовершенствовано с помощью учёта различных типов связей между словами [3]. В работе [4] описан улучшенный Y. Xie подход, использующий для обработки различных видов оборотов семантические правила, соответствующие различным способам выражения тональности. Этот подход достигает F_1-мер 0.76 и 0.68 на наборах данных комментариев из Facebook и твитов соответственно. В работе O. Appel [7] к этому подходу были добавлены новые правила и механизм автоматизированного расширения тонального словаря с помощью машинного обучения. Авторы подхода исследовали качество его работы на трёх наборах данных. На наборах данных комментариев из Facebook и твитов была достигнута доля правильных ответов (accuracy) около 0.88. На третьем наборе данных, состоящем из отзывов на фильмы, доля правильных ответов составила 0.76.

Предыдущая работа авторов настоящего исследования [1] посвящена адаптации подхода O. Appel к русскому языку. Наибольшую сложность при адаптации составило то, что в исходной работе семантические правила были реализованы как шаблоны над предложением, представленным в виде списка слов, что требовало соблюдения строгого порядка слов, тогда как в русском языке порядок слов в предложении строго не регламентирован. Для решения этой проблемы правила были реализованы в виде алгоритмов над деревом синтаксических связей (dependency parse tree), отражающим зависимости между словами в предложении. Эксперименты с адаптированным подходом на корпусе отзывов на отели показали возможность достичь с использованием подхода, основанного на правилах, результатов, сравнимых с наилучшими для русского языка, достигаемыми с использованием нейронных сетей.

Дерево разбора предложения — представление иерархической структуры слов в предложении [8]. Существуют два подхода к анализу структуры предложения с помощью деревьев разбора. Первый рассматривает предложение как иерархию слов, упорядоченных в соответствии с синтаксическими связями. Узлы дерева синтаксических связей — слова; рёбра показывают синтаксическое согласование между ними. Другой подход основан на членении фраз на их составные части. Листьями дерева составляющих являются слова, ветвями — фразы, а рёбра отражают связь типа «целое — часть».

Деревья составляющих и семантические правила успешно применяются для анализа смысла во множестве приложений, например, в распознавании вопросов и команд [9, 10] и выявлении логической связности предложений [11]. Применительно к анализу тональности существует набор данных Stanford Sentiment Treebank (SST) [12], включающий маркировку отдельных частей предложений по тональности для 215154 фраз из 11855 предложений. Большинство методов анализа тональности, задействующих деревья составляющих, используют нейронные сети и набор данных SST для обучения и оценки качества. Существует два основных подхода. Первый — Tree-LSTM [13, 14], архитектура нейронной сети, упорядочивающая узлы LSTM в соответствии с деревом разбора так, что каждый из них соответствует узлу дерева разбора. Второй подход обрабатывает предложение как систему фраз, каждая из которых может, в свою очередь, также делиться на отдельные фразы; тональность же каждой из фраз может быть определена по тональностям её составляющих. Одной из последних моделей, основанных на данной идее, является нейронная сеть SentiBERT, определяющая тональность фразы с помощью представления в виде трансформеров её составных частей и машинного обучения [15].

Адаптация вышеперечисленных подходов, основанных на нейронных сетях, к русскому языку в настоящее время чрезвычайно затруднительна, поскольку требует большого корпуса, схожего с SST, то есть имеющего разметку по тональности каждого узла каждого из деревьев составляющих. Следовательно, разработка основанных на правилах подходов, не требующих больших корпусов для обучения, представляется для русского языка перспективным направлением исследований.

2. Описание алгоритма

Структура фраз предложения описывает, как из слов формируются высказывания [8]. Для обработки структуры фраз в данной работе используется дерево составляющих. Листьями дерева составляющих являются слова, ветвями — фразы, а рёбра отражают связь типа «целое — часть». Корнем дерева составляющих является предикативное ядро предложения.

В данной работе используются деревья составляющих, а не деревья синтаксических связей (отражающие синтаксическое согласование между словами), потому что дерево составляющих лучше отражает структуру фраз предложения. Следовательно, семантические правила оказывается проще построить с использованием деревьев составляющих. Поскольку для русского языка не существует анализаторов, строящих деревья составляющих, но есть анализаторы для построения деревьев синтаксических связей, был разработан вспомогательный алгоритм, преобразующий дерево синтаксических связей в дерево составляющих. Этот алгоритм основан на группировке слов, являющихся потомками одного слова, во фразы в соответствии с частеречной разметкой и типами синтаксических связей.

Алгоритм определения тональности, предлагаемый в данной статье, основан на предположении, что тональность фразы может быть определена на основе тональностей её составляющих. Был составлен набор правил, описывающих, как вывести тональность фразы по тональностям её частей, которые определяются с помощью рекурсивного применения того же набора правил. Тональности отдельных слов, к которым не может быть применён рекурсивный вызов, определяются с помощью тонального словаря RuSentiLex-2017 [16] (с изменениями, описанными в предыдущей работе авторов [1]). Несмотря на низкое качество этого словаря [1], для русского языка не существует лучших альтернатив.

Более формально: пусть N — узел дерева составляющих, $C(N)$ — множество дочерних узлов N, а $S(N)$ — тональность N. Алгоритм начинает свою работу от корня дерева составляющих N_r, который соответствует предикативному ядру обрабатываемого предложения; таким образом, $S(N_r)$ — тональность предложения.

Алгоритм вычисляет $S(N)$ для заданного узла N следующим образом:

- если $C(N) = \varnothing$, т. е. узел представляет одиночное слово, то $S(N)$ — тональность N, определённая в соответствии со словарём;
- иначе рекурсивно вычислить $S(N_c)$ для каждого $N_c \in C(N)$; выбрать подходящее семантическое правило в зависимости от типа $C(N)$ (см. ниже) и вычислить по нему $S(N)$.

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

Первая группа относится к фразам, состоящим из однородных членов предложения, например, *ум, честь и совесть* (положительная тональность), *и талантливый, и ленивый* (нейтральная тональность), *разработанный коллективно и обречённый на неудачу* (отрицательная тональность), и т. д. 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 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.

Table 1. Modifier words and phrases **Таблица 1.** Слова- и фразы-модификаторы

Модификация	Слова и фразы
делают нейтральную или положительную тональность отрицательной; делают отрицательную тональность положительной	не, нет, прекратить, разрушить, перестать, избегать, противоречить, уменьшать, утратить, ни один
делают отрицательную тональность нейтральной; остальную оставляют неизменной	понимать, принимать
делают нейтральную тональность положительной; остальную оставляют неизменной	способность, обеспечить, позволить, выполнить, помочь, увеличить, достичь, доказать, надеяться, очень, также
делают положительную тональность нейтральной; остальную оставляют неизменной	крайне
делают любую тональность нейтральной	вроде бы, несмотря на

Во второй группе правил одна часть фразы модифицирует тональность другой части фразы. Например, рассмотрим фразы *объём производства увеличивается* и *дефицит бюджета увеличивается*. В первой из них слово *увеличивается* изменяет нейтральную тональность части *объём производства* и придаёт всей фразе положительную тональность. Во второй то же самое слово *увеличивается* никак не изменяет отрицательную тональность фразы *дефицит бюджета*; тональность всей фразы оказывается отрицательной. По результатам обсуждения с экспертами-лингвистами были выделены 26 наиболее употребительных русских слов-модификаторов и фраз-модификаторов (таблица 1).

В третьей группе тональность фразы является результатом 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. Например, имеющая отрицательную тональность фраза *отчислить лучшего студента* состоит из сказуемого *отчислить*, имеющего отрицательную тональность, и дополнения *лучший студент*, тональность которого положительна. В то же время, нейтральная фраза *заслуженно отчислили* состоит из сказуемого *отчислили*, имеющего отрицательную тональность, и обстоятельства *заслуженно*, имеющего положительную тональность.

После обсуждения с экспертами-лингвистами были сформулированы правила, описывающие типичные результаты соединения тональностей для всех возможных комбинаций членов предложения на русском языке (таблица 2).

Правило, которое применяется к данной фразе, выбирается следующим образом:

- если фраза состоит из однородных членов, то применяется правило для однородных членов;
- если одна из частей фразы является словом-модификатором или фразой-модификатором, то применяется соответствующее им правило;
- иначе применяется правило для соединения частей фразы.

Следует отметить, что используемая в описанном алгоритме модель имеет ряд ограничений. Первое относится к ситуациям, когда тональность фразы не может быть определена только по тональностям её составляющих. Например, рассмотрим фразу *после проведения акции сайт организаторов исчез*. Несмотря на то, что её части *сайт организаторов* и *исчез после проведения акции* нейтральны, фраза в целом имеет отрицательную тональность. Второе ограничение связано с тем, что предложенные правила достаточно просты и не покрывают все случаи определения тональности фразы по тональностям её частей. Например, фраза *мастерски обманывает* содержит имеющее отрицательную тональность сказуемое *обманывает* и положительное обстоятельство *мастерски*. Соответствующее правило говорит о том, что такая фраза нейтральна, тогда как на самом деле её тональность отрицательна.

Table 2. Sentiment of combination of sentence parts

Таблица 2. Тональность комбинаций членов предложения

Первая часть фразы	Вторая часть фразы	Тональность первой части фразы	Тональность второй части фразы	Тональность фразы
Подлежащее	Сказуемое	Положительная	Положительная	Положительная
			Нейтральная	Положительная
			Отрицательная	Отрицательная
		Нейтральная	Положительная	Положительная
			Нейтральная	Нейтральная
			Отрицательная	Отрицательная
		Отрицательная	Положительная	Нейтральная
			Нейтральная	Отрицательная
			Отрицательная	Отрицательная
	Определение	Положительная	Положительная	Положительная
			Нейтральная	Положительная
			Отрицательная	Отрицательная
		Нейтральная	Положительная	Положительная
			Нейтральная	Нейтральная
			Отрицательная	Отрицательная
		Отрицательная	Положительная	Отрицательная
			Нейтральная	Отрицательная
			Отрицательная	Отрицательная
	Дополнение	Положительная	Положительная	Положительная
			Нейтральная	Положительная
			Отрицательная	Нейтральная
		Нейтральная	Положительная	Нейтральная
			Нейтральная	Нейтральная
			Отрицательная	Отрицательная
		Отрицательная	Положительная	Нейтральная
			Нейтральная	Отрицательная
			Отрицательная	Отрицательная
Сказуемое	Обстоятельство	Положительная	Положительная	Положительная
			Нейтральная	Положительная
			Отрицательная	Отрицательная
		Нейтральная	Положительная	Положительная
			Нейтральная	Нейтральная
			Отрицательная	Отрицательная
		Отрицательная	Положительная	Нейтральная
			Нейтральная	Отрицательная
			Отрицательная	Отрицательная
	Дополнение	Положительная	Положительная	Положительная
			Нейтральная	Положительная
			Отрицательная	Отрицательная
		Нейтральная	Положительная	Нейтральная
			Нейтральная	Нейтральная
			Отрицательная	Отрицательная
		Отрицательная	Положительная	Отрицательная
			Нейтральная	Отрицательная
			Отрицательная	Отрицательная
Дополнение	Определение	Положительная	Положительная	Положительная
			Нейтральная	Нейтральная
			Отрицательная	Отрицательная
		Нейтральная	Положительная	Положительная
			Нейтральная	Нейтральная
			Отрицательная	Отрицательная
		Отрицательная	Положительная	Нейтральная
			Нейтральная	Отрицательная
			Отрицательная	Отрицательная

Несмотря на эти ограничения, представляется разумным провести эксперименты даже с такой простой моделью, чтобы оценить её потенциал для определения тональности, а также провести анализ ошибок, на основе которого можно будет улучшить модель в будущем.

3. Корпуса, использованные для экспериментов

Для оценки способности предложенного алгоритма определять тональности достаточно простых предложений были проведены эксперименты на корпусе отзывов на отели, использованном в предыдущем исследовании авторов [1]. В большинстве предложений этого корпуса тональность выражена чётко и определённо, поэтому задача определения тональности оказывается достаточно простой.

Для оценки применимости и эффективности предложенного алгоритма в более сложных случаях были проведены эксперименты на корпусе, состоящем из предложений, относящихся к нескольким предметным областям, и содержащем существенную долю сложных предложений.

В свободном доступе имеется нескольких корпусов русскоязычных предложений с разметкой по тональности: SentiRuEval-2015 Subtask 2, SentiRuEval2016, RuTweetCorp, LINIS Crowd, Kaggle Russian News Dataset и RuReviews [5]. К сожалению, ни один из них не подходит для данного исследования: некоторые содержат тексты, а не предложения (Kaggle Russian News Dataset, RuReviews), в других слишком много ошибок разметки (RuTweetCorp, LINIS Crowd) или нет достаточного количества сложных предложений (SentiRuEval). Поэтому был создан OpenSentimentCorpus, основанный на OpenCorpora — открытом корпусе предложений из новостных и публицистических текстов на русском языке, относящихся к различным предметным областям.

Чтобы увеличить долю сложных предложений и отбросить слишком простые, малополезные для оценки качества, из корпуса были исключены предложения короче семи слов. Все оставшиеся были размечены как минимум тремя экспертами.

Каждое предложение было отмечено как *положительное*, *нейтральное* или *отрицательное*, если оно выражает соответствующую тональность; как *смешанное*, если оно выражает одновременно и положительную, и отрицательную тональность, например, *Пелевин пишет не в пример хуже Акунина, но у Пелевина есть что сказать читателю*; либо как *сомнительное*, если его тональность непонятна, оно содержит нечёткую авторскую позицию, а также во всех остальных случаях неуверенности эксперта.

Результаты разметки обрабатывались следующим образом:

- Предложения, получившие как минимум одну отметку *сомнительное* или одновременно отметки *положительное* и *отрицательное*, были исключены из рассмотрения как имеющие нечётко выраженную или непонятную тональность.
- Окончательная тональность предложения определялась в соответствии с преобладающей среди всех экспертов отметкой. Если таковой не находилось (например, если только 2 из 3 отметок были положительными, а одна — нейтральной), то такое предложение исключалось.

Построенный OpenSentimentCorpus — корпус предложений, относящихся к различным предметным областям, подходящий для экспериментов по классификации на 2, 3 или 4 класса. Поскольку рассматриваемый в настоящей работе алгоритм определяет предложения лишь положительной, отрицательной или нейтральной тональности, предложения смешанной тональности были исключены из дальнейшего рассмотрения. Объём OpenSentimentCorpus достаточен для получения достоверных результатов экспериментов.

Распределения предложений по классам для обоих рассматриваемых корпусов приведены в таблице 3.

В обоих корпусах предложения распределены по классам неравномерно. Примерно половина предложений корпуса отзывов на отели имеет положительную тональность, четверть — отрицательную, и только одна пятая — нейтральную. Главная причина этого лежит в особенностях предметной

Table 3. Classes of sentences in hotel reviews corpus and OpenSentimentCorpus

Таблица 3. Классы предложений с различной тональностью в корпусах отзывов на отели и OpenSentimentCorpus

Корпус	Корпус отзывов на отели	OpenSentimentCorpus
Положительная	639	536
Нейтральная	232	2441
Отрицательная	333	1510
Всего	1204	4487

области — большинство авторов отзывов выражает свою удовлетворённость или неудовлетворённость, а не приводит факты. В OpenSentimentCorpus, напротив, большая часть предложений содержит нейтральные высказывания, не выражающие мнения их авторов, примерно треть предложений имеет отрицательную тональность, и только десятая часть — положительную. Тем не менее, OpenSentimentCorpus содержит более 500 положительных предложений, чего достаточно для оценки качества работы алгоритма.

OpenSentimentCorpus доступен онлайн и может быть загружен по адресу: https://github.com/yarfruct/open-sentiment-corpus.

4. Эксперименты

Качество работы предложенного в данной работе рекурсивного алгоритма оценивалось на корпусе отзывов на отели и на OpenSentimentCorpus и сравнивалось с качеством работы основанного на правилах подхода, адаптированного для русского языка в [1].

Как метрики качества использовались простое и взвешенное среднее точности, полноты и F_1-меры. Простое среднее (также называемое арифметическим средним) — сумма значений соответствующих метрик для всех классов, делённая на количество классов. Взвешенное среднее — сумма метрик качества для классов, умноженных на количество предложений в каждом классе, и делённая на общее количество предложений в корпусе. Причиной использования взвешенного среднего является дисбаланс классов в корпусах, который может привести к неправильной оценке качества работы алгоритма.

Метрики качества и матрицы ошибок для корпуса отзывов на отели приведены в таблицах 4 и 5. Предложенный рекурсивный алгоритм определяет положительную и отрицательную тональность на данном корпусе достаточно точно, но несколько хуже адаптированного подхода. Это снижение в основном обусловлено увеличившейся долей предложений с положительной тональностью, неверно классифицированных как нейтральные. Хуже всего и рекурсивный алгоритм, и адаптированный подход разделяют предложения с отрицательной и нейтральной тональностями, хотя рекурсивный алгоритм и делает это лучше адаптированного подхода.

Table 4. Sentiment classification performances of the proposed recursive algorithm and the algorithm adapted in [1] on the hotel reviews corpus

Таблица 4. Качество работы предложенного рекурсивного алгоритма и адаптированного в работе [1] подхода на корпусе отзывов на отели

Алгоритм	Рекурсивный алгоритм			Адаптированный подход			Количество
Класс предложений	Точность	Полнота	F_1-мера	Точность	Полнота	F_1-мера	предложений
Положительный	0.89	0.77	0.82	0.88	0.88	0.88	639
Нейтральный	0.43	0.75	0.55	0.48	0.75	0.58	232
Отрицательный	0.86	0.64	0.73	0.94	0.57	0.71	333
Среднее	0.73	0.72	0.70	0.77	0.73	0.73	1204
Взвешенное среднее	0.79	0.73	0.75	0.82	0.77	0.78	1204

Доля правильных ответов рекурсивного алгоритма = 0.73
Доля правильных ответов адаптированного подхода = 0.77

Table 5. Sentiment classification confusion matrices of the proposed recursive algorithm and the rule-based approach adapted in [1] on the hotel reviews corpus

Таблица 5. Матрицы ошибок предложенного рекурсивного алгоритма и адаптированного в работе [1] подхода на корпусе отзывов на отели

Тональность	Рекурсивный алгоритм			Адаптированный подход			Всего
предсказ. / реальн.	Положит.	Нейтр.	Отрицат.	Положит.	Нейтр.	Отрицат.	
Положительная	491	137	11	546	72	3	639
Нейтральная	33	175	24	50	173	9	232
Отрицательная	29	92	212	24	118	191	333

Метрики качества и матрицы ошибок на корпусе OpenSentimentCorpus приведены в таблицах 6 и 7. Качество работы рекурсивного алгоритма на OpenSentimentCorpus в среднем примерно на 5–6 % ниже, чем на корпусе отзывов на отели. Наиболее заметны сложности с определением положительной тональности: алгоритм корректно классифицировал только половину имеющих её предложений, а существенную их долю, наоборот, классифицировал как имеющие отрицательную тональность. Напротив, отрицательную тональность алгоритм определяет на OpenSentimentCorpus лучше, чем на корпусе отзывов на отели, а также точнее разделяет отрицательную и нейтральную тональности. В целом, рекурсивный алгоритм работает на OpenSentimentCorpus примерно на 7 % лучше, чем адаптированный подход. Единственное существенное преимущество адаптированного подхода — меньшее количество неверно классифицированных предложений с положительной тональностью.

Table 6. Sentiment classification performances of the proposed recursive algorithm and the adapted rule-based approach on OpenSentimentCorpus

Таблица 6. Качество работы предложенного рекурсивного алгоритма и адаптированного в работе [1] подхода на корпусе OpenSentimentCorpus

Алгоритм	Рекурсивный алгоритм			Адаптированный подход			Количество
Класс предложений	Точность	Полнота	F_1-мера	Точность	Полнота	F_1-мера	предложений
Положительный	0.45	0.49	0.47	0.30	0.71	0.42	536
Нейтральный	0.76	0.74	0.75	0.72	0.65	0.68	2441
Отрицательный	0.70	0.70	0.70	0.77	0.52	0.62	1510
Среднее	0.63	0.64	0.64	0.60	0.62	0.57	4487
Взвешенное среднее	0.70	0.70	0.70	0.68	0.61	0.63	4487

Доля правильных ответов рекурсивного алгоритма = 0.70
Доля правильных ответов адаптированного подхода = 0.61

Table 7. Sentiment classification confusion matrices of the proposed recursive algorithm and the adapted rule-based approach on OpenSentimentCorpus

Таблица 7. Матрицы ошибок предложенного рекурсивного алгоритма и адаптированного в работе [1] подхода на корпусе OpenSentimentCorpus

Тональность	Рекурсивный алгоритм			Адаптированный подход			Всего
предсказ. / реальн.	Положит.	Нейтр.	Отрицат.	Положит.	Нейтр.	Отрицат.	
Положительная	261	209	66	378	131	27	536
Нейтральная	244	1800	397	640	1596	205	2441
Отрицательная	81	371	1058	228	504	778	1510

5. Анализ ошибок

Для выяснения причин снижения эффективности работы алгоритма на OpenSentimentCorpus была собрана информация о 150 неверно классифицированных предложениях (по 50 предложений из каждого класса). Они были разделены на 4 группы (таблица 8) в зависимости от причины неверной классификации:

- некорректное построение синтаксического дерева;
- некорректное определение тональности одиночных слов;
- несовершенство правил;
- тональность определяется высокоуровневой структурой предложения.

Table 8. Error groups of the proposed recursive algorithm on OpenSentimentCorpus

Таблица 8. Группы ошибок предложенного алгоритма на корпусе OpenSentimentCorpus

Ошибка	% положит. предложений	% нейтр. предложений	% отрицат. предложений
Некорректное построение синтаксического дерева	12	4	24
Некорректное определение тональности одиночных слов	24	12	34
Несовершенство правил	26	22	2
Тональность определяется высокоуровневой структурой предложения	38	62	40

Некорректное построение синтаксического дерева включает в себя ошибки определения частей речи, построения дерева составляющих, лемматизации и все остальные, порождённые синтаксическим анализатором. Для предложений положительной и нейтральной тональностей это наименее частые ошибки, тогда как почти четверть ошибок классификации предложений с отрицательной тональностью вызвана несовершенством синтаксического анализатора. Причиной этого может быть более сложная структура предложений с отрицательной тональностью по сравнению с остальными.

Ошибки, связанные с неверным определением тональности одиночных слов, возникают, когда словарная тональность слова не соответствует реальной. Главная и наиболее важная их причина — несовершенство тонального словаря. В словаре RuSentiLex отсутствуют многие слова, имеющие ярко выраженную тональную окраску, такие как *больно*, *истерика*, *ломать*, *взаимовыгодный*, *позабавить*, *благодаря*. Ошибки этой группы также возникают из-за слов, тональность которых меняется в зависимости от контекста и предметной области. Например, слово *исторический* обычно нейтрально, но в контексте значимости некоторого события оно приобретает положительную окраску, например, *исторический момент*. Следует отметить, что RuSentiLex-2017 предоставляет некоторую информацию об омонимии, но для того, чтобы использовать её для улучшения алгоритма, необходимо отдельное исследование. Ошибки, причиной которых является неверное определение тональности одиночных слов, влияют в основном на тональные предложения, и только десятая часть ошибок классификации нейтральных предложений относится к этой группе. Возможная причина такого дисбаланса — меньшее количество тональных слов в нейтральных предложениях по сравнению с тональными.

Последние две группы ошибок связаны с ограничениями алгоритма. Как отмечалось выше, используемые им семантические правила довольно просты и не всегда правильно определяют тональность конкретной фразы. Например, нейтральное предложение *Я правда не видела нового шестого фильма* неверно классифицировано как имеющее отрицательную тональность из-за того, что правило обработки «не» определяет тональность фразы *не видела нового шестого фильма* как отрицательную, как следствие, тональность всего предложения также определяется как отрицательная. Лишь

Table 9. Accuracy of modifier words and phrases on OpenSentimentCorpus

Таблица 9. Доля случаев правильного применения слов- и фраз-модификаторов на корпусе OpenSentimentCorpus

Модификация	Слова и фразы	Доля, %
Делают нейтральную или положительную тональность отрицательной; делают отрицательную тональность положительной	не, нет, прекратить, разрушить, перестать, избегать, противоречить, уменьшать, утратить, ни один	56
Делают отрицательную тональность нейтральной; остальную оставляют неизменной	понимать, принимать	76
Делают нейтральную тональность положительной; остальную оставляют неизменной	способность, обеспечить, позволить, выполнить, помочь, увеличить, достичь, доказать, надеяться, очень, также	77
Делают положительную тональность нейтральной; остальную оставляют неизменной	крайне	100
Делают любую тональность нейтральной	вроде бы, несмотря на	80

небольшое количество предложений с отрицательной тональностью классифицируется неверно из-за ошибок этой группы, в отличие от предложений с положительной и нейтральной тональностью.

Вероятно, оценки алгоритма несколько смещены в сторону отрицательной тональности: он верно определяет тональность большинства отрицательных фраз, но также считает отрицательными многие фразы, которые таковыми не являются. Анализ правильности использования слов- и фраз-модификаторов (таблица 9) показывает, что они выбраны достаточно адекватно.

Последняя группа ошибок включает предложения, неверно классифицированные из-за того, что тональность некоторых фраз не может быть определена на основе тональностей их составных частей, а определяется высокоуровневой структурой предложения. Например, рассмотрим предложение *С возрастом детям становится всё более интересен и интернет: его регулярно посещают 15 % младших школьников и почти 60 % подростков.* Оно содержит положительно окрашенную информацию о росте использования интернета, но общий контекст и стиль речи показывают, что автор не выражает собственное мнение, а просто констатирует факт, следовательно, предложение является нейтральным. Такие ошибки чаще всего возникают при определении тональности нейтральных предложений, вероятно, из-за того, что существенная их часть извлечена из новостных сообщений, которые могут содержать тонально окрашенные факты, но общий их контекст нейтрален.

Подводя итоги анализа ошибок, можно сделать следующие выводы. Ошибки различных групп распределены по классам тональности неравномерно: те из них, которые связаны с некорректным синтаксическим анализом, приводят к некорректной классификации предложений с отрицательной тональностью чаще, чем всех остальных; корректное определение тональности отдельных слов наиболее существенно для классификации предложений с положительной и отрицательной тональностью; несовершенство правил влияет на качество классификации предложений с положительной и нейтральной тональностью в большей степени по сравнению с отрицательной; недостаток анализа высокоуровневой структуры предложений влияет на все классы тональности, но сильнее всего — на нейтральный.

Заключение

В данной статье предложен рекурсивный алгоритм определения тональности русскоязычных предложений, основанный на семантических правилах, и проведена оценка качества его работы. Для улучшения анализа структуры предложений были использованы деревья составляющих и предложен набор семантических правил, позволяющих определить тональность фразы на основе тональностей её частей. Качество работы предложенного алгоритма было экспериментально

оценено и сопоставлено с качеством работы другого основанного на правилах подхода, адаптированного к русскому языку в предыдущей работе авторов, на двух корпусах: корпусе отзывов на отели и OpenSentimentCorpus, состоящим из достаточно сложных предложений, относящихся к различным предметным областям и извлечённых из новостных текстов. Эксперименты показали, что на корпусе отзывов на отели предложенный рекурсивный алгоритм показывает несколько худшие по сравнению с адаптированным подходом результаты: взвешенные F_1-меры составили 0.75 и 0.78 соответственно. На OpenSentimentCorpus качество работы обоих методов оказалось значительно ниже, чем на корпусе отзывов на отели (что было ожидаемо ввиду сложности предложений и более нетривиальных способов выражения тональности), однако предложенный алгоритм работает значительно лучше, чем адаптированный подход: F_1-меры составили 0.70 и 0.63 соответственно. Это свидетельствует о том, что в сложных случаях предложенный алгоритм имеет существенное преимущество над адаптированным подходом.

Несмотря на то, что в среднем эффективность алгоритма на OpenSentimentCorpus не слишком высока, следует иметь в виду, что набор семантических правил, использованных в экспериментах, достаточно простой, а корпус по своим характеристикам значительно отличается от обычно используемых в исследованиях по анализу тональности на русском языке. Следует также отметить, что предложения с отрицательной тональностью определяются достаточно точно, что может свидетельствовать о соответствии правил реально существующим способам выражения отрицательной тональности.

Будущие исследования могут быть связаны с усовершенствованием алгоритма для повышения качества определения положительной и нейтральной тональности. На основе анализа ошибок этой цели предположительно можно достичь за счёт повышения точности определения тональности одиночных слов (в том числе с помощью расширения используемого тонального словаря) и добавления новых, более точных правил, отражающих различные способы выражения тональности.

References

[1] I. Paramonov and A. Poletaev, “Adaptation of Semantic Rule-Based Sentiment Analysis Approach for Russian Language”, in *Proceedings of 30th Conference of Open Innovations Association FRUCT*, 2021, pp. 155–164.

[2] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis”, in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005, pp. 347–354.

[3] L. K.-W. Tan, J.-C. Na, Y.-L. Theng, and K. Chang, “Sentence-level sentiment polarity classification using a linguistic approach”, in *International Conference on Asian Digital Libraries*, 2011, pp. 77–87.

[4] Y. Xie, Z. Chen, K. Zhang, Y. Cheng, D. K. Honbo, A. Agrawal, and A. N. Choudhary, “MuSES: multilingual sentiment elicitation system for social media data”, *IEEE Intelligent Systems*, vol. 29, no. 4, pp. 34–42, 2014.

[5] 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.

[6] M. A. M. Shaikh, H. Prendinger, and M. Ishizuka, “Sentiment assessment of text by analyzing linguistic features and contextual valence assignment”, *Applied Artificial Intelligence*, vol. 22, no. 6, pp. 558–601, 2008.

[7] O. Appel, F. Chiclana, J. Carter, and H. Fujita, “A hybrid approach to the sentiment analysis problem at the sentence level”, *Knowledge-Based Systems*, vol. 108, pp. 110–124, 2016.

[8] S. Kahane and N. Mazziotta, “Syntactic Polygraphs. A Formalism Extending Both Constituency and Dependency”, in *Proceedings of the 14th Meeting on the Mathematics of Language*, 2015, pp. 152–164.

[9] Y. Gao, J.-G. Lou, and D. Zhang, *A Hybrid Semantic Parsing Approach for Tabular Data Analysis*, 2019. arXiv: 1910.10363v2 [cs.AI].

[10] J. Li, H. Tan, and M. Bansal, "Improving Cross-Modal Alignment in Vision Language Navigation via Syntactic Information", in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1041–1050.

[11] Z. Marji, A. Nighojkar, and J. Licato, "Probing the Natural Language Inference Task with Automated Reasoning Tools", in *The Thirty-Third International Flairs Conference*, 2020.

[12] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank", in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

[13] K. S. Tai, R. Socher, and C. D. Manning, "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks", in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015.

[14] Y. Zhang and Y. Zhang, "Tree communication models for sentiment analysis", in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 3518–3527.

[15] D. Yin, T. Meng, and K.-W. Chang, "SentiBERT: A Transferable Transformer-Based Architecture for Compositional Sentiment Semantics", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3695–3706.

[16] N. V. Loukachevitch and A. V. Levchick, "Creating a General Russian Sentiment Lexicon", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 1171–1176.