УДК 004.048:519.767.6
DOI 10.17513/snt.40727

РАЗРАБОТКА МОДЕЛИ И ЧИСЛЕННЫХ МЕТОДОВ ДЛЯ ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ТЕКСТОВ НА ОСНОВЕ СИСТЕМЫ С ПРЕДОПРЕДЕЛЕННОЙ СЕМАНТИКОЙ

Ванюлин А. Н., Алексеева Н. Р., Давыдова О. В.

Федеральное государственное бюджетное образовательное учреждение высшего образования «Чувашский государственный университет имени И. Н. Ульянова», Чебоксары, Российская Федерация, e-mail: alexis-04@mail.ru

Целью исследования является разработка и апробация математической модели и численных методов для определения тональности текстов на основе системы с предопределенной семантикой. В качестве материалов исследования использовалась выборка из 1000 пользовательских отзывов. Для обучения системы была проведена ручная разметка тональности на уровне отдельных предложений, где каждому предложению присваивалась метка: +1 (позитивное), 0 (нейтральное) или -1 (негативное). Ключевым в предлагаемом методе является вычисление семантического спектра для языковых единиц (слов, фраз, предложений). Этот спектр представляет собой числовой вектор, который формируется по специальному алгоритму, учитывающему не только набор символов в слове, но и их порядок, причем последние символы вносят наибольший вклад в результирующий спектр. Для учета структуры предложения и минимизации влияния порядка слов был разработан алгоритм построения древовидной структуры предложения. На этапе обучения система формировала три базы данных (для позитивных, нейтральных и негативных слов), куда записывались скорректированные семантические спектры слов и информация об их наиболее вероятном уровне в дереве. Результаты тестирования метода на выборке, не участвовавшей в обучении, показали такую же точность распознавания, что и с помощью метода Bag of Words, что продемонстрировало сопоставимую точность. Основной вывод исследования заключается в том, что предложенный метод позволяет достигать уровня точности, сравнимого со стандартными методами, без привлечения внешних лингвистических ресурсов, таких как базы данных словоформ.

Ключевые слова: математическое моделирование, численные методы, система с предопределенной семантикой, семантический спектр, анализ тональности, комплекс программ, классификация текстов

DEVELOPMENT OF A MODEL AND NUMERICAL METHODS FOR DETERMINING THE SENTIMENT OF TEXTS BASED ON A SYSTEM WITH PREDEFINITIVE SEMANTICS

Vanyulin A. N., Alekseeva N. R., Davydova O. V.

Federal State Budgetary Educational Institution of Higher Education "Chuvash State University named after I. N. Ulyanov", Cheboksary, Russian Federation, e-mail: alexis-04@mail.ru

The aim of the study is to develop and test a mathematical model and numerical methods for determining the sentiment of texts based on a system with predefined semantics. A sample of 1,000 user reviews was used as research materials. To train the system, sentiment was manually annotated at the sentence level, with each sentence assigned a label: +1 (positive), 0 (neutral), or -1 (negative). The key to the proposed method is the calculation of a semantic spectrum for linguistic units (words, phrases, sentences). This spectrum is a numerical vector generated using a special algorithm that takes into account not only the set of characters in a word but also their order, with the last characters making the largest contribution to the resulting spectrum. To account for sentence structure and minimize the influence of word order, an algorithm for constructing a tree-like sentence structure was developed. During the training phase, the system created three databases (for positive, neutral, and negative words), which contained the adjusted semantic spectra of words and information about their most probable level in the tree. Testing the method on a sample not included in the training phase demonstrated the same recognition accuracy as the Bag of Words method, demonstrating comparable precision. The main conclusion of the study is that the proposed method achieves a level of accuracy comparable to standard methods without relying on external linguistic resources, such as word form databases.

Keywords: mathematical modeling, numerical methods, system with predefined semantics, semantic spectrum, sentiment analysis, software suite, text classification

Введение

Автоматическое распознавание тональности текстов применяется во многих сферах деятельности, это могут быть, например, маркетинговые и политологические исследования, анализ новостей, поддержание обратной связи с пользователями сайтов и т. д. В то же время задача определения тональности текстов является одной из разновидностей задач их классификации [1–3].

В случае определения тональности тексты обычно делятся на три группы:

– тексты с отрицательной тональностью;
– тексты нейтральной тональности;
– тексты с положительной тональностью.

Для автоматического определения тональности используются самые разнообразные методы, начиная с самых простых (по набору ключевых слов) [4] и заканчивая самыми современными на основе методов машинного обучения [5, 6]. На актуальность данной темы указывает и достаточно большой объем публикаций по этому направлению.

Цель исследования – разработка и апробация математической модели и численных методов для определения тональности текстов на основе системы с предопределенной семантикой (СПС).

Материал и методы исследования

1. Математическая модель семантического спектра

В работах [7, 8] описаны основные особенности СПС и предложена специфическая функция для определения семантики слов, учитывающая не только символы, входящие в слово, но порядок их вхождения в слово. Результатом определения этой функции является некоторый набор чисел, который можно интерпретировать как семантический спектр слова.

Общая схема формирования спектра состоит в последовательном присоединении спектра очередного символа слова к уже сформированному спектру, состоящему из спектров предыдущих символов. Формально для каждого последующего символа применяется формула

$$s_0 = \frac{s_1 + \frac{s_2}{2}}{2}, \qquad (*)$$

где s_0 – вектор формируемой семантики;
s_1 – семантика предшествующего символа;
s_2 – семантика очередного символа.

Длина спектра равна количеству символов, имеющихся на клавиатуре компьютера, а значение каждого элемента спектра определяется количеством вхождений символа и порядком его нахождения в слове.

Предложенная функция может быть также интерпретирована как спектр «ощущений» системы каждого конкретного слова, то есть в качестве «органов чувств» системы выступают отдельные символы текста. Именно в таком контексте получаемым спектрам присвоено значение «семантические». В этом контексте СПС представляет собой совершенно оригинальный вариант моделирования процессов мышления.

Особенностью получаемых спектров является то, что величина сигнала в спектре зависит от положения символа в слове – то есть наименьший сигнал имеет первый символ, а наибольший – последний. В качестве примеров на рис. 1 приведены спектры слов «кот» и «ток».

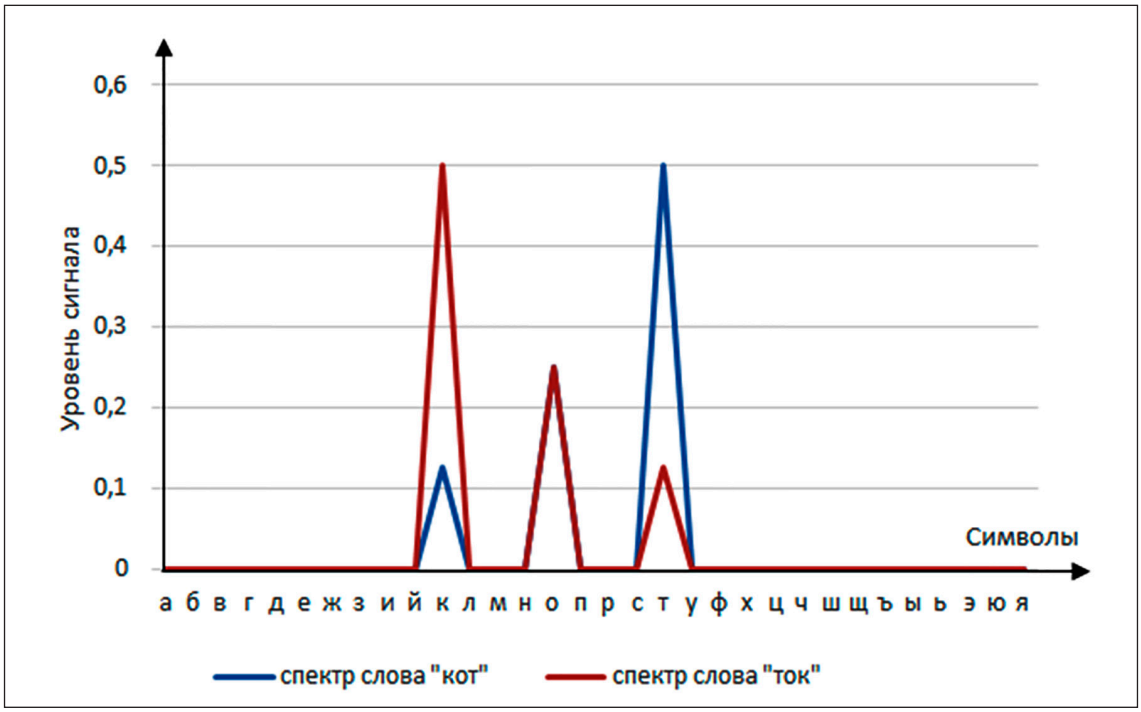

Рис. 1. Семантические спектры слов «кот» и «ток»
Примечание: составлен авторами по результатам данного исследования

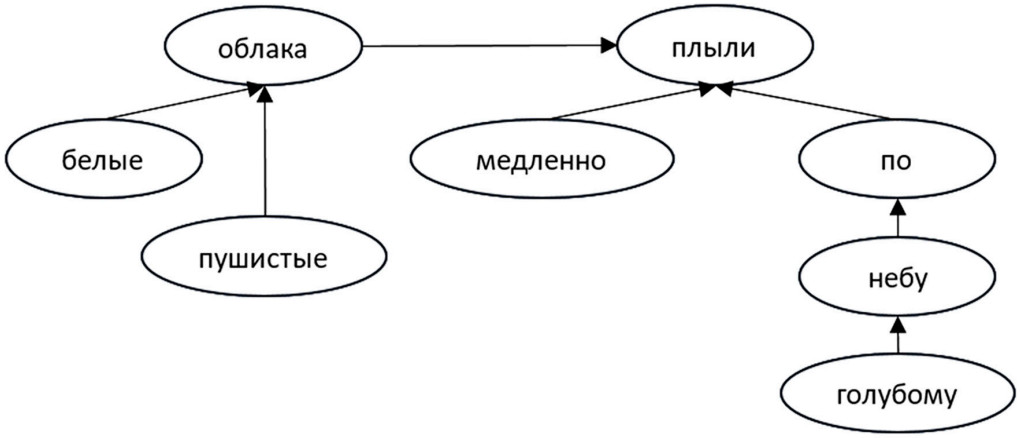

Рис. 2. Структура фразы «белые пушистые облака медленно плыли по голубому небу»
Примечание: составлен авторами по результатам данного исследования

Для получения семантических спектров фраз была применена та же технология, но в качестве исходных компонентов использованы спектры слов, из которых состоит фраза. На вид итогового спектра фразы также влияет порядок вовлекаемых в формирование спектра слов. В случае использования того же порядка, что и для отдельных слов, вклад начальных слов в значения элементов спектра будет заметно меньше, чем вклад конечных слов.

Аналогично можно сформировать спектр всего отзыва, где в качестве исходных компонентов должны быть использованы спектры составляющих его фраз.

2. Численные методы обработки текста

Для минимизации влияния линейного порядка слов на итоговый спектр предложения разработан численный метод построения древовидной структуры предложения.

Согласно источнику[1], любое предложение можно представить в виде некоторой древовидной структуры. Например, для фразы «Белые пушистые облака медленно плыли по голубому небу» соответствующая структура может быть такой, как на рис. 2.

Для автоматического построения такого дерева используется алгоритм, основанный на корреляции семантических спектров слов с эталонными спектрами тональности.

Отметим, что для построения деревьев нецелесообразно использовать комбинаторные методы генерации деревьев, поскольку такие методы имеют факториальную зависимость количества деревьев от количества объектов. Поэтому для автоматического построения дерева разработан алгоритм, основанный на корреляции семантических спектров слов с эталонными спектрами тональности.

Алгоритм построения дерева предложения:

1. Для каждого слова предложения вычисляется коэффициент корреляции его семантического спектра с эталонными спектрами позитивной, нейтральной и негативной тональности (вид эталонов приведен на рис. 3).

2. Слова сортируются по убыванию коэффициента корреляции.

3. Диапазон значений коэффициентов разбивается на фиксированное число уровней (в работе принято 6 уровней). Каждому слову присваивается номер уровня, соответствующий интервалу, в который попадает его коэффициент.

4. Построение дерева осуществляется по правилам: слова одного уровня не объединяются; слово нижнего уровня объединяется со словом верхнего уровня справа, а при его отсутствии – с левым словом верхнего уровня. Результатом объединения является формирование семантического спектра вышестоящего узла.

Существенным моментом алгоритма является использование эталонных спектров. Формально эталонные спектры можно сформировать на основе конкретных слов. Например, спектры слов «позитивный», «нейтральный» и «негативный» можно было бы использовать как эталонные для определения позитивности, нейтральности или негативности предложения.

[1] Русский язык: учебник для студентов учреждений высшего профессионального образования / Л. Л. Касаткин, Е. В. Клобуков, Л. П. Крысин и др.; под ред. Л. Л. Касаткина. 4-е изд., перераб. М.: Издательский центр «Академия», 2011. 780 с. [Электронный ресурс]. URL: https://academia-moscow.ru/ftp_share/_books/fragments/fragment_14850.pdf (дата обращения: 23.03.2026). ISBN 978-5-7695-7997-4.

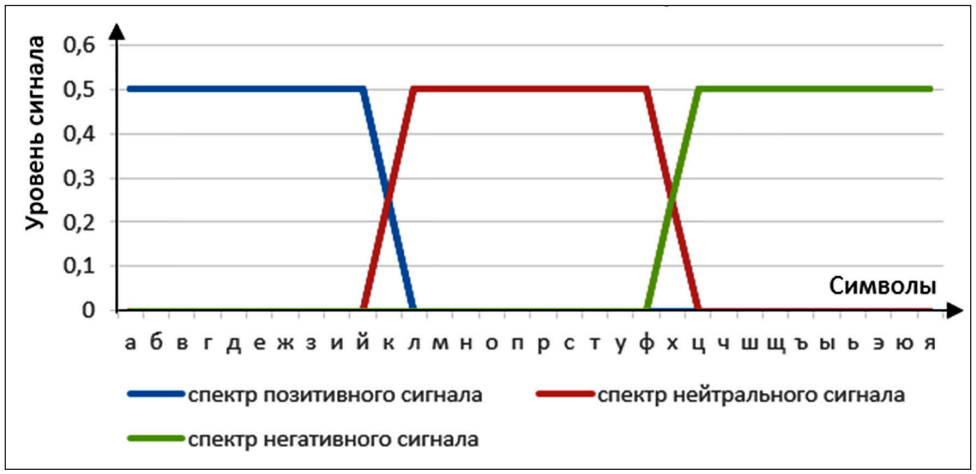

Рис. 3. Вид эталонных спектров
Примечание: составлен авторами по результатам данного исследования

Однако указанные слова имеют одинаковое окончание, что из-за указанной выше особенности формирования спектров (последние символы вносят основной вклад в вид спектра) приводит к мало чем отличающимся спектрам. Перебор других слов-кандидатов на роль слов с эталонными спектрами также не приводит к получению сильно отличающихся спектров. Поэтому в качестве эталонных приняты спектры, представленные на рис. 3.

Последующие эксперименты показали, что именно такой подход к выбору эталонов позволяет наиболее точно осуществлять дифференциацию предложений по тональности.

3. Обучение модели и классификация

Непосредственная работа с данными производилась по стандартной схеме машинного обучения [9, 10] и состояла из двух этапов – обучение системы и последующее использование обученной системы непосредственно для распознавания [11]. Для обучения и тестирования использовалась выборка из 1000 отзывов.

Содержанием первого этапа являлась ручная разметка обучающей выборки отзывов, полученной на основе случайного разделения. Объем предназначенного для обучения (разметки) текста определялся эмпирически и увеличивался до получения устойчивой статистики распознавания, что составило 80 % от общего количества обработанных отзывов. Тексты отзывов были взяты с сайта[2].

Обработанные отзывы как текстовые документы имеют следующую статистику:

– среднее количество предложений в отзыве – 7,41 предложений на один отзыв;

– максимальное количество предложений в отзыве 28;

– среднее количество слов в предложении – 9,86;

– максимальное количество слов в предложении – 74;

– средняя длина слова в символах – 7,45;

– слово максимальной длины – 24 символа (слово – «клиентоориентированность»).

Входящие в состав одного отзыва предложения могут иметь все три вида тональности (положительную, нейтральную и отрицательную). Поэтому фактически при разметке приходилось оценивать тональность каждого предложения, входящего в отзыв. При этом разметка заключалась в том, что каждому предложению в отзыве вручную ставилась метка: позитивная (1), нейтральная (0) или негативная (-1). В рамках данного исследования общая тональность всего отзыва определялась как набор чисел, равных количеству позитивных, негативных и нейтральных предложений в отзыве (или их процентном соотношении).

Необходимо отметить, что для некоторых предложений расстановка меток носила достаточно вариативный и субъективный характер. В самом простом случае метка ставилась под влиянием общего контекста отзыва, так как в отрыве от контекста само по себе предложение могло быть нейтральным, но под влиянием контекста ему могла быть поставлена позитивная или негативная метка.

Пример такого отзыва:

«*Весной 2025 получил заказ: саженцы плодовых растений, всего 40 штук разных наименований на приличную сумму. Из них 3 саженца были абсолютно сухими*».

[2] Отзывы об интернет-магазинах. [Электронный ресурс]. URL: https://otzyvshops.com (дата обращения: 23.03.2026).

Второму предложению была поставлена метка «негативное», хотя в отрыве от контекста это совершенно нейтральное предложение. Отметим, что в рамках используемых в системе методов совершенно нереально ожидать от системы понимания того, что «*сухие саженцы*» – это плохо.

В некоторых же случаях в одном и том же предложении были и отрицательные, и положительные интонации, и тогда расстановка меток носила полностью субъективный характер.

Пример такого отзыва:

«*Заказал пару сабо с хорошими скидками, отправили быстро, но пришла пара с браком*».

Для настройки параметров модели использовалась обучающая выборка, состоящая из предложений с известной тональностью (разметка +1, 0, -1). В процессе обучения для каждого слова вычислялся его семантический спектр, который корректируется путем усреднения по всем вхождениям данного слова в предложения одной тональности. Для каждого слова также фиксировался наиболее вероятный уровень в дереве (по частоте встречаемости). Полученные данные записывались в три базы данных (позитивную, нейтральную и негативную), которые представляют собой параметры обученной модели.

Классификация нового предложения выполняется следующим образом:

– предложение разбивается на слова;

– для каждого слова из трех баз данных извлекаются его усредненный спектр и наиболее вероятный уровень;

– по описанному выше алгоритму строится дерево предложения и вычисляется его итоговый семантический спектр;

– полученный спектр сравнивается с эталонными спектрами (рис. 3) с помощью коэффициента корреляции;

– предложению присваивается тональность, соответствующая эталону с максимальной корреляцией.

4. Программный комплекс

Все численные методы реализованы в виде комплекса программ на языке VBA в среде Microsoft Excel. Программа включает модули для вычисления семантических спектров, построения деревьев, обучения и классификации. Комплекс зарегистрирован в Федеральной службе по интеллектуальной собственности (свидетельство № 2020666193).

Результаты исследования и их обсуждение

В качестве примера приводятся некоторые промежуточные результаты работы системы. Например, фраза «Первый заказ очень долго оформляла» имеет негативную метку, и по результатам сравнения с эталоном для каждого слова получены следующие уровни для соответствующих слов (табл. 1).

В табл. 2 представлен порядок объединения семантики слов в предложении.

На рис. 4 показано дерево, соответствующее полученным переходам.

Сравнение структур, приведенных на рис. 4 и на рис. 2, показывает, что разработанный алгоритм позволяет генерировать достаточно корректные синтаксические структуры.

В работе [12] представлены результаты оценки тональности текстов с помощью стандартного метода компьютерной лингвистики – метода Bag of Words.

Основные результаты определения тональности на тестирующей выборке приведены в табл. 3. На главной диагонали таблицы приведено количество предложений с правильно определенной тональностью, остальные значения – количество ошибочных определений.

Таблица 1

Уровни слов для фразы «Первый заказ очень долго оформляла»

Уровень дерева	Слова					Контекст
	1	2	3	4	5	
	первый	заказ	очень	долго	оформляла	
0	0	0	0	0	0	**1**
1	0	0	0	0	**1**	0
2	0	0	0	0	0	0
3	0	**1**	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	**1**	0	**1**	**1**	0	0

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Таблица 2

Порядок объединения семантики слов в предложении
«Первый заказ очень долго оформляла»

Номер перехода	1	2	3	4	5
Начальное слово	1	3	4	2	5
Конечное слово	2	5	5	5	Контекст

Примечание: составлена авторами на основе полученных данных в ходе исследования.

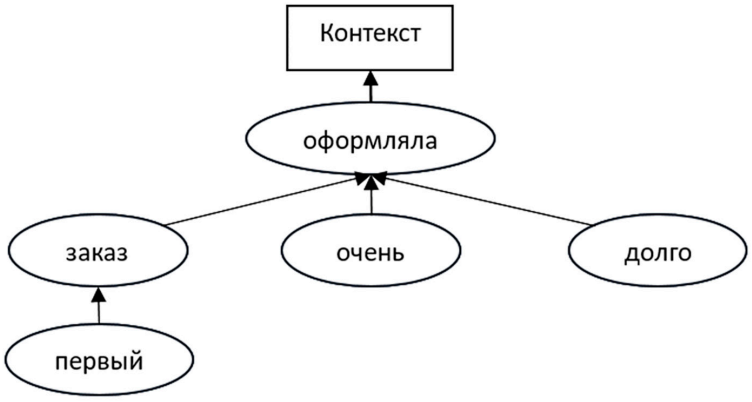

Рис. 4. Структура фразы «Первый заказ очень долго оформляла» в виде дерева отношений
Примечание: составлен авторами по результатам данного исследования

Таблица 3

Численные результаты распознавания тональности предложений
на тестирующей выборке

Результат распознавания	Реальная тональность предложений		
	Позитивная	Нейтральная	Негативная
Позитивная	109	31	34
Нейтральная	22	141	40
Негативная	25	37	177
Всего предложений	156	209	251

Примечание: составлена авторами на основе полученных данных в ходе исследования

В табл. 4 приведены результаты расчетов метрик, наиболее часто применяемых для оценки качества машинного обучения[3].

Данные табл. 4 показывают, что точность определения (Precision) с помощью предлагаемого метода сопоставима с точностью, получаемой с помощью стандартных методов компьютерной лингвистики [13–15]. Значения остальных метрик показывают, что предложенный метод работает стабильно для всех трех классов, а разброс между метриками Precision и Recall не превышает 0,1, что указывает на отсутствие существенного дисбаланса в распознавании.

Заключение

В результате исследования разработана математическая модель представления текста в виде семантических спектров, основанная на рекуррентной формуле (*).

Предложены и реализованы следующие численные методы:

– вычисления спектров слов и фраз;

– построения древовидной структуры предложения, снижающего вычислительную сложность по сравнению с полным перебором;

– классификации текстов по тональности на основе корреляционного анализа.

[3] Михайличенко А. А. Аналитический обзор методов оценки качества алгоритмов классификации в задачах машинного обучения // Ежеквартальный рецензируемый, реферируемый научный журнал «Вестник АГУ». 2022. Вып. 4 (311). С. 52–59. DOI: 10.53598/2410-3225-2022-4-311-52-59.

Таблица 4

Результаты расчетов метрик качества распознавания

Метрика качества	Тональность		
	Позитивная	Нейтральная	Негативная
Precision	0,698	0,675	0,705
Recall	0,626	0,695	0,741
F-метрика	0,661	0,684	0,722

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Созданный программный комплекс (свидетельство о регистрации № 2020666193) позволяет автоматизировать процесс определения тональности без использования внешних лингвистических баз данных.

Основные преимущества предложенного подхода:

– отсутствует необходимость в использовании внешних лингвистических баз данных, таких как, например, базы данных основных словоформ, синонимов и т. д.;

– возможность дальнейшего развития за счет учета контекста;

– возможность обработки текстов на любом языке, поскольку метод работает на уровне символов.

Полученная точность определения сопоставима с известными методами, что подтверждает перспективность дальнейших исследований в направлении совершенствования численных методов на основе семантических спектров.

Список литературы

1. Богданова Т. Ф., Бойчук Е. И. Основные теоретические проблемы, связанные с понятием тональности текста // Верхневолжский филологический вестник. 2020. № 4 (23). С. 136–141. DOI: 10.20323/2499-9679-2020-4-23-136-141.

2. Конурбаев М. Э. Тембр как тензор: многомерная модель жанрово-стилистического анализа литературного текста // Russian Linguistic Bulletin. 2026. № 1 (73). URL: https://rulb.org/archive/1-73-2026-january/10.60797/RULB.2026.73.2 (дата обращения: 22.03.2026). DOI: 10.60797/RULB.2026.73.2. EDN: SSNECB.

3. Двойникова А. А., Карпов А. А. Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных // Информационно-управляющие системы. 2020. № 4. С. 20–30. DOI: 10.31799/1684-8853-2020-4-20-30.

4. Рубцова Ю. В. Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015. № 1 (109). С. 72–78. DOI: 10.15827/0236-235X.109.072-078.

5. Райимкулов А. Б., Ашералиева М. Ж., Корякина Ю. С. Методы и технологии измерения тональности текста // Проблемы автоматики и управления. 2025. № 3 (54). С. 101–109. URL: https://pau.imash.kg/index.php/pau/ru/article/view/546 (дата обращения: 23.03.2026).

6. Максименко О. И., Беляков М. В. Сентимент-анализ как инструмент лингвоэмотиологии: оценка потенциала систем анализа тональности текста // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2025. Т. 16. № 3. С. 760–782. DOI: 10.22363/2313-2299-2025-16-3-760-782.

7. Ванюлин А. Н., Алексеева Н. Р., Мочалова Т. А. Лингвистические основы алгоритмов компьютерной обработки текстов на основе систем с предопределенной семантикой // Современные наукоемкие технологии. 2020. № 3. С. 35–39. URL: http://www.top-technologies.ru/ru/article/view?id=37936 (дата обращения: 23.03.2026). DOI: 10.17513/snt.37936.

8. Ванюлин А. Н., Алексеева Н. Р. Алгоритмы реализации систем с предопределенной семантикой на основе концепции семантических полей // Современные наукоемкие технологии. 2022. № 5-1. С. 7–11. URL: https://top-technologies.ru/ru/article/view?id=39142 (дата обращения: 22.03.2026). DOI: 10.17513/snt.39142.

9. Васильев В. В. Парадигмы анализа тональности и сентимент-анализа региональных интернет-СМИ на примере новостных порталов Якутии // Филология: научные исследования. 2025. № 12. С. 335–345. URL: https://nbpublish.com/library_read_article.php?id=77208 (дата обращения: 23.03.2026). DOI: 10.7256/2454-0749.2025.12.77208.

10. Жаксыбаев Д. О., Мизамова Г. Н. Алгоритмы обработки естественного языка для понимания семантики текста // Труды ИСП РАН. 2022. Т. 34. № 1. С. 135–150. DOI: 10.15514/ISPRAS-2022-34(1)-10.

11. Самигулин Т. Р., Джурабаев А. Э. У. Анализ тональности текста методами машинного обучения // Научный результат. Информационные технологии. 2021. Т. 6. № 1. С. 55–62. DOI: 10.18413/2518-1092-2021-6-1-0-7.

12. Ванюлин А. Н., Алексеева Н. Р. Определение тональности текстов методами компьютерной лингвистики: анализ отзывов о торговых сетях // Современные наукоемкие технологии. 2025. № 5. С. 27–31. URL: https://top-technologies.ru/ru/article/view?id=40386 (дата обращения: 23.03.2026). DOI: 10.17513/snt.40386.

13. Плешакова Е. С., Гатауллин С. Т., Осипов А. В., Романова Е. В., Самбуров Н. С. Эффективная классификация текстов на естественном языке и определение тональности речи с использованием выбранных методов машинного обучения // Вопросы безопасности. 2022. № 4. С. 1–14. URL: https://nbpublish.com/library_read_article.php?id=38658 (дата обращения: 22.03.2026). DOI: 10.25136/2409-7543.2022.4.38658.

14. Хорошилов А. А., Козловская Я. Д., Мусабаев Р. Р., Красовицкий А. М., Хорошилов А. А. Определение тональности сообщений СМИ методами их концептуального анализа // Моделирование и анализ данных. 2019. № 4. С. 67–79. DOI: 10.17759/mda.2019090405.

15. Басина П. А., Дунаева Д. О., Саркисова А. Ю. Валидация моделей машинного обучения для автоматизированного определения тональности русскоязычных текстов // Вестник Томского государственного университета. 2022. № 485. С. 206–216. DOI: 10.17223/15617793/485/23.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: The authors declare that there is no conflict of interest.

Финансирование: Авторы заявляют об отсутствии внешнего финансирования.

Financing: The research was performed without external funding.