Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

«Обзор методов анализа тональности естественно-языковых текстов»

Студент: ИУ7-53Б _____ М. Д. Маслова
(группа) (подпись, дата) (И. О. Фамилия)

Руководитель: _____ А. А. Оленев
(подпись, дата) (И. О. Фамилия)

2021 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой ИУ-7
_____ И. В. Рудаков
« _____ » _____ 20 ____ г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме Обзор методов анализа тональности естественно-языковых текстов

Студент группы ИУ7-53Б

Маслова Марина Дмитриевна
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)
учебная

Источник тематики (кафедра, предприятие, НИР) НИР

График выполнения НИР: 25% к 4 нед., 50% к 7 нед., 75% к 11 нед., 100% к 14 нед.

Техническое задание

Описать существующие методы анализа тональности естественно-языковых текстов. Выделить критерии оценки описанных методов. Классифицировать и сравнить существующие решения по выделенным критериям.

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на 15-25 листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)
Презентация на 8-10 слайдах.

Дата выдачи задания « 03 » сентября 2021 г.

Руководитель НИР _____ А. А. Оленев
(Подпись, дата) (И.О.Фамилия)

Студент _____ М. Д. Маслова
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Современное развитие сети Интернет позволяет пользователям ежедневно создавать и выкладывать в открытый доступ различную информацию, в связи с чем происходит накопление большого числа данных, одной из наиболее распространенных форм хранения которых являются тексты на естественном языке. Необходимость анализа накопленных массивов текстовых данных привела к развитию направления обработки естественного языка (NLP — Natural Language Processing), одной из основных задач которого стал анализ тональности или сентимент-анализ, заключающийся в выделении из текстов субъективных мнений и эмоций. Востребованность анализа тональности во многих областях и невозможность ручной обработки большого числа текстов послужили разработке и развитию многочисленных автоматических методов, решающих задачу сентимент-анализа [1].

Целью данной работы является классификация методов анализа тональности естественно-языковых текстов.

Для достижения поставленной цели решаются следующие задачи:

– рассматриваются основные подходы к анализу тональности;

– описываются методы анализа тональности, относящиеся к каждому из подходов;

– предлагаются критерии оценки качества описанных методов;

– сравниваются методы по предложенным критериям оценки;

– выделяются методы, показывающие лучшие результаты по одному или нескольким критериям.

1 Анализ предметной области

В данном разделе обоснована актуальность задачи, представлены основные определения и формализация задачи.

1.1 Актуальность задачи

В современном мире огромную роль в жизни каждого человека играет Интернет. Люди общаются в социальных сетях, ведут блоги, оставляют отзывы о товарах, услугах, фильмах, книгах и т. п. За счет этого в открытом доступе находится огромный объем данных, который позволяет проводить точные анализы для решения каких-либо задач.

Большая часть накопленных данных представлена в виде текстовой информации, поэтому становится актуальной задача анализа текстов на естественном языке [2]. Одной из этих задач является анализ тональности или сентимент-анализ. За счет того, что такой анализ может быть проведен для текста, написанного на любую тему, его применение возможно во многих сферах:

– мониторинг общественного мнения относительно товаров и услуг, в том числе в режиме реального времени, с целью определения их достоинств и недостатков с точки зрения покупателей и улучшения их характеристик [3];

– анализ политических и социальных взглядов пользователей (например, влияние мер, предпринятых для борьбы с вирусом COVID-19, на жизнь людей);

– исследование рынка и прогнозирование цен на акции [4];

– выявление случаев эмоционального насилия и пресечение противоправных действий [5].

Решение описанных задач требует анализа большого количества текстов, что делает невозможной их ручную обработку. Также при оценке тональности текста человеком трудно соблюсти критерии этой оценки [3]. Таким образом, возникает необходимость в автоматизированных системах анализа.

При этом в отличие от традиционной обработки текста в анализе тональности незначительные вариации между двумя элементами текста существенно меняют смысл (например, добавление частицы «не»). Обработку естественного языка затрудняет обильное использования носителями средств вырази-

тельности и переносных значений слов и фраз. Также основной из проблем сентимент-анализа является разная окраска одного и того же слова в текстах на различные тематики: слово, которое считается положительным в одной, в то же время считается отрицательным в другой [4].

С учетом широкого применения анализ тональности и описанных сложностей, возникает необходимость в формализации поставленной проблемы и разработки методов для ее решения.

1.2 Основные определения

Анализ тональности текста (или *сентимент-анализ*) — область компьютерной лингвистики, ориентированная на извлечение из текстов субъективных мнений и эмоций. *Тональность* — это мнение, отношение и эмоции автора по отношению к объекту, о котором говорится в тексте. Чаще всего под задачей анализа тональности текста понимают определение текста к одному из двух классов: «положительный» или «отрицательный». В некоторых случаях добавляют третий класс «нейтральных» текстов [1].

В настоящее время выделяют три основных подхода к определению тональности текста [1]:

– *лингвистический подход* предполагает анализ лексики в тексте на основе заранее созданных словарей, правил и шаблонов;

– *подход, основанный на машинном обучении*, строится на обучении и автоматическом построении классифицирующей функции на основе некоторых данных, полученных из текстов, тональность которых известна;

– *гибридный подход* сочетает в себе подходы как на основе словарей, правил и шаблонов, так и на основе машинного обучения;

Несмотря на различные подходы, можно выделить несколько этапов анализа тональности текста [6]:

– предварительная обработка текста;

– извлечение информативных признаков или векторизация текста;

– классификация;

– оценка результата работы.

Первым этапом анализа тональности является предварительная обработка текста, которая также происходит в несколько этапов:

– приведение текста к *единому регистру* для сокращения количества слов, которые необходимо хранить одновременно;

– *удаление пунктуации и шума* (упоминаний пользователей, ссылок, хештегов), *исправление опечаток и ошибок* [7];

– *токенизация* или разбиение исходного текста на лексемы, в простейшем случае — разбиение по пробельным символам [8];

– *удаление стоп-слов*, то есть слов не несущих никакой смысловой нагрузки, с целью повышения точности;

– *стемминг* или *лемматизация* — приемы приведения слов форм слова к общему виду; в случае стемминга происходит получение корня слова путем отбрасывания приставок, суффиксов и окончаний, в случае лемматизации — воспроизводится начальная форма слова, то есть та форма, которая представлена в словаре [7];

– обработка отрицаний [9].

На втором этапе анализа происходит представление текста в виде вектора. Для этого используется один из многочисленных методов, таких как «мешок слов» (Bag of Words, представляет текст в виде неупорядоченного набора слов), Word2Vec (нейронная сеть, генерирующая векторы слов) и др. Этап векторизации текста для методов лингвистического подхода не является обязательным, так как в данном случае происходит работа непосредственно с текстами, а не с их векторами [6].

На третьем этапе происходит собственно определение тональности текста с использованием одного из методов, описанных ниже.

На последнем этапе производится оценка результатов работы классификатора.

1.3 Формализация задачи

В данной работе ставится задача анализа методов определения принадлежности заданного естественно-языкового текста к одному из двух классов:

– положительный;

– отрицательный.

При этом определяется лишь факт принадлежности тому или иному классу, и оценка вероятности отношения текста к каждому классу не проводится.

2 Описание существующих решений

В данном разделе представлено краткое описание существующих методов анализа тональности естественно-языковых текстов.

2.1 Лингвистический подход

Методы использующие лингвистический подход можно разделить на две основные категории:

– методы на основе правил;

– методы на основе лексики.

Последние, в свою очередь, основываются либо на словарях слов, либо на корпусах текстов.

2.1.1 Методы, основанные на правилах

Методы, основанные на правилах, (или *rule-based methods*) для определения класса используют большой набор созданных в ручную правил вида «если $\rightarrow$ то» [6]. Левая часть каждого правила показывает набор признаков, а правая часть — метку класса [10].

В случае анализа тональности набор признаков может быть представлен словом, словосочетанием или другой более сложной языковой конструкцией, а каждый класс — обозначен необходимым образом, например, символ «+» — положительный, «−» — отрицательный. В таком случае одними из простейших правил могут быть описаны формулами (2.1), (2.2):

$$\{\text{хороший}\} \rightarrow \{+\} \tag{2.1}$$

$$\{\text{плохой}\} \rightarrow \{-\} \tag{2.2}$$

Большой набор правил описанного вида позволяет осуществлять прогнозирования тональности анализируемого текста [11].

Данные алгоритмы имеют высокую точность в узких областях тем текстов, однако их обобщение на более широкий круг тем затруднительно. Также процесс создания необходимых правил является трудоемким за счет их определения человеком, а не компьютером [12].

В целях ускорения процесса разработки для создания набора правил может использоваться машинное обучение, поэтому в некоторых научных работах [11] [13] данные методы относят к методам машинного обучения.

2.1.2 Методы, основанные на тональных словарях

Методы, основанные на тональных словарях, (или *dictionary-based methods*) относят к методам, основанным на лексике.

В данном подходе вручную собирается небольшой набор уникальных слов и словосочетаний, важных для тематики анализируемых текстов, с их тональными оценками (весами) [1]. Полученный набор увеличивается путем поиска в онлайн-словарях синонимов и антонимов к каждому элементу, входящему в набор. Найденные слова и словосочетания добавляются в исходный набор, и операция повторяется до тех пор, пока при очередном повторении не будет получено ни одного нового элемента набора [14]. Итоговый набор, полученный описанным методом, образует тональный словарь.

При анализе текста в итоговом наборе осуществляется поиск каждого слова и выбор его веса. Если слова нет в словаре, то считают, что оно нейтрально, и присваивают ему вес, равный нулю. По найденным весам высчитывается принадлежности анализируемого текста к тому или иному классу тональности [1].

Данный подход показывает высокую точность для конкретных областей, но полученный в результате поиска синонимов и антонимов словарь может быть корректно использован только для той предметной области, для которой он составлялся. При этом выделение начального небольшого набора слов и словосочетаний требует хороших знаний в анализируемой предметной области, а следовательно, возникает необходимость в изучении данной области или поиске специалиста, что приводит к дополнительным трудозатратам [1].

2.1.3 Методы, основанные на корпусах

Методы, основанные на корпусах, (или *corpus-based methods*) также относят к методам, основанным на лексике.

Данный подход так же, как и предыдущий, начинает работу с небольшого набора слов, однако, в отличие от методов, основанных на тональных

словарях, методы, основанные на корпусах, осуществляют поиск слов для расширения исходного списка в большом наборе анализируемых текстов, а не в онлайн-словарях [10]. Таким образом учитывается не только самостоятельная тональная оценка слова, но и контекст, в котором это слово находится. При этом существует два способа для учета контекста [11]:

– статистический подход учитывает число появлений слова в положительных и отрицательных текстах, при преобладании в первых считается, что тональная оценка слова положительна, при преобладании во вторых — отрицательна, при совпадении частоты появлений — нейтральна; при этом учитывается тот факт, что два слова часто встречающиеся вместе в одном контексте с высокой вероятностью будут иметь одинаковый знак оценки [15];

– семантический подход опирается на различные принципы вычисления сходства между словами и присваивает словам близким по смыслу одинаковые значения веса [15].

Достоинством этого подхода является рассмотрение слова ни как самостоятельной единицы с фиксированным весом, а как части единого текста с тональной оценкой, зависящей от окружения [10]; однако, так же как и в случае методов на основе тональных словарей, список слов полученный для одного набора текстов не может быть использован для анализа текстов иной предметной области, так как одно и то же слово в различных областях может вносить разный вес эмоциональной окраски [1].

2.2 Методы машинного обучения

В подходе к анализу тональности с точки зрения машинного обучения набор текстов разделяется на две различные сборки: для обучающей и тестовой выборки. Далее для контролируемого (с учителем) обучения тексты из обучающей выборки размечаются, то есть указывается их класс тональности, а для неконтролируемого (без учителя) обучения дополнительных данных не указывается [16].

2.2.1 Наивный байесовский классификатор

Наивный байесовский классификатор [17] является контролируемым вероятностным подходом, который определят тональность текста путем нахож-

дения наиболее вероятного класса C_T для данного текста T, что может быть описано формулой (2.3):

$$C_T = \underset{C}{\operatorname{argmax}} P(C|T), \tag{2.3}$$

где $P(C|T)$ — вероятность того, что текст T принадлежит классу C, которая может быть определена по формуле Баейса (2.4).

$$P(C|T) = \frac{P(C)P(T|C)}{P(T)}, \tag{2.4}$$

где $P(C)$ — вероятность того, что встретится класс C независимо от анализируемого текста;

$P(T|C)$ — вероятность встретить текст T среди текстов класса C;

$P(T)$ — вероятность того, что встретится текст T среди всех текстов.

Вероятность $P(T)$ в формуле (2.4) не влияет на отношение текста к тому или иному классу, поэтому может быть опущена.

При представлении текста в виде вектора входящих в него слов и предположении об их независимости (что и дало данному классификатору название «наивный») вероятность встретить текст T среди текстов класса C может быть вычислена по формуле (2.5):

$$P(T|C) = P(w_1, w_2, ..., w_N|C) = \\ = P(w_1|C) \cdot P(w_2|C) \cdot ... \cdot P(w_N|C) = \prod_{i=1}^{N} P(w_i|C), \tag{2.5}$$

где w_i — i-ое слово в тексте T;

$P(w_i|C)$ — вероятность встретить i-ое слово в классе C;

N — количество слов в тексте.

Таким образом, с учетом формул (2.3)-(2.5) итоговый класс анализируемого текста может быть найден по формуле (2.6):

$$C_T = \underset{C}{\operatorname{argmax}} P(C)P(T|C) = \underset{C}{\operatorname{argmax}} P(C) \prod_{i=1}^{N} P(w_i|C). \tag{2.6}$$

Наивный байесовский классификатор прост в понимании, но в нем делает-

ся предположение о независимости признаков, которое в естественно-языковых текстах обычно не подтверждается. Однако, несмотря на всю простоту и ограничение на независимость, наивный байесовский классификатор может показывать высокую точность при анализе тональности текста [1].

2.2.2 Логистическая регрессия

Логистическая регрессия является методом линейного классификатора, использующим для прогнозирования вероятности принадлежности текстов к классу путем вычисления значения логистической функции $f(z)$, описывающейся формулой (2.7):

$$f(z) = \frac{1}{1 + e^{-z}}. \tag{2.7}$$

Параметр логистической функции z описывается формулой (2.8):

$$z = \theta^T x = \theta_0 + \theta_1 x_1 + ... + \theta_n x_n, \tag{2.8}$$

где x — вектор-столбец, в случае анализа тональности описывающий данный текст;

θ — вектор-столбец коэффициентов, получаемых в ходе обработке обучающей выборки.

Для определения класса текста, он представляется в виде вектора-столбца x, далее вычисляется значение логистической функции. Исходя из графика логистической функции, представленного на рисунке 2.1, если полученное значение больше 0.5 текст считается положительным, иначе отрицательным [18].

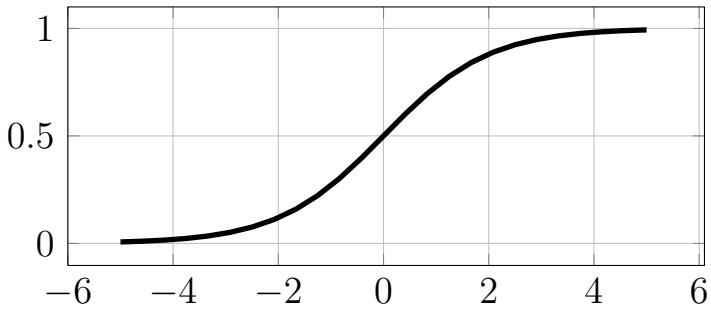

Рисунок 2.1 – График логистической функции

Логистическая регрессия может показывать высокую точность, однако для этого необходима качественная предобработка и отбор признаков для представления текста в виде вектора-столбца [1].

2.2.3 Метод максимума энтропии

Метод максимума энтропии [19] является вероятностным классификатором, который основан на принципе максимальной энтропии. По данному принципу распределения вероятности являются равномерными (имеют максимальную энтропию), если нет оснований считать иначе, то есть предположения о независимости слов, как в случае наивного байесовского классификатора, не делается, а при обучении максимизируются их веса с помощью итерационной процедуры.

Вероятность $P(C|T)$ того, что текст T принадлежит классу C, в данном случае определяется формулой (2.9):

$$P(C|T) = \frac{1}{Z(T)} \exp\Big(\sum_i \lambda_i f_i(T, C)\Big), \tag{2.9}$$

где $Z(T)$ — коэффициент нормализации, гарантирующий выполнение условия нормировки, вычисляющийся по формуле (2.10);

λ_i — вес i-ого признака;

$f_i(T, C)$ — функция принадлежности i-ого признака текста T классу C.

$$Z(T) = \sum_C \exp\Big(\sum_i \lambda_i f_i(T, C)\Big), \tag{2.10}$$

где обозначения соответствуют обозначениям в формуле (2.9).

В силу отсутствия предположения о независимости признаков как в случае наивного байесовского классификатора метод максимума энтропии позволяет добиться лучших результатов с сохранением простоты реализации и необходимости для обучения малого числа данных [1].

2.2.4 *k*-ближайших соседей

k-ближайших соседей — метод, работа которого заключается в поиске обучающих текстов, наиболее похожих на анализируемый, при этом построение

обучающих данных происходит с учетом соотношений текстов друг с другом.

Для определения тональности текста с помощью данного метода вычисляется расстояние между тестовыми данными и уже обработанными обучающими. Чаще всего в качестве расстояния используют косинусное сходство (формула (2.11)), соответствующее косинусу угла θ между векторами $\vec{A}$ и $\vec{B}$, которые задают сравниваемые тексты.

$$\cos\theta = \frac{\vec{A}\cdot\vec{B}}{\|\vec{A}\|\|\vec{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\cdot\sqrt{\sum\limits_{i=1}^{n} B_i^2}} \tag{2.11}$$

После вычисления всех расстояний в их наборе ищутся k наименьших, причем k определяется заранее. И в конце анализируемому тексту сопоставляется тот класс, к которому относится большинство из k выбранных соседей.

Данный метод прост в реализации, однако имеет большое время выполнения в силу необходимости полного перебора [20].

2.2.5 Дерево решений

Классификатор *дерева решений* строит обучающие данные в древовидную структуру: выбирается слово, тексты, которые его содержат, помещаются на правую ветвь дерева, остальные — в левую; для каждой ветви процедура повторяется до тех пор, пока листья не будут содержать определенное минимальное количество записей, которые используются для определения тональности. Таким образом, внутренние узлы дерева представляют условие, являющееся проверкой на наличие или отсутствие одного или нескольких слов, а листья содержат либо минимальный набор текстов, по которым можно определить тональность анализируемого, либо метку класса, к которому будет принадлежать текст, удовлетворяющий всем условиям, включенным в путь от корня к данному листу.

При анализе тональности текста, не входящего в обучающую выборку, для него, начиная с корня построенного дерева, проверяются условия во внутренних узлах для поиска необходимого листа, по информации из которого определяется тональность [14].

Метод дерева решений является рекурсивным, прост в реализации, требу-

ет минимальной предварительной обработки данных и дает высокую точность при решении задачи анализа тональности текста на большом количестве данных, однако данный метод при небольших изменениях в обучающей выборке, получает кардинально разные результаты на тестовых данных [1].

2.2.6 Случайный лес

Для решения проблемы дерева решений с зависимостью от обучающих данных используется ансамбль решающих деревьев, или *случайный лес*. В данном методе строится большое количество решающих деревьев на разных обучающих данных по алгоритму, описанному в предыдущем пункте. Для определения тональности текста, он анализируется с помощью каждого дерева, класс, выбранный большинством деревьев, определяет тональность анализируемого текста.

Данный метод решает проблемы одиночного решающего дерева, однако требует больших временных затрат на построение деревьев [1].

2.2.7 Метод опорных векторов

Метод опорных векторов является контролируемым линейным подходом. При решении задачи классификации текстов, в том числе на положительные и отрицательные, тексты обучающей выборки представляются в виде векторов, которым соответствуют точки в n-мерном пространстве, где n — размерность вектора. Цель метода заключается в поиске такой гиперплоскости пространства, что сумма расстояний от ближайших к этой гиперплоскости точек в каждом классе (называющихся опорными векторами) максимальна (рисунок 2.2) [11].

Данный метод показывает высокую точность при определении тональности текстов, так как способен к масштабированию и может работать с большим количеством признаков на больших выборках [1].

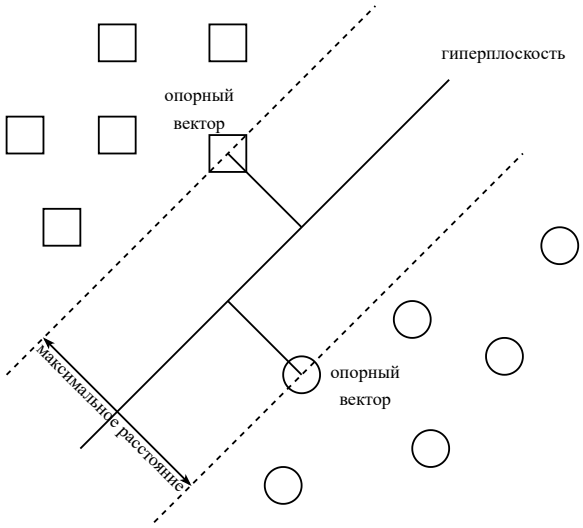

Рисунок 2.2 – Визуализация метода опорных векторов

2.2.8 Нейронные сети

Нейронные сети [1] являются математической моделью, построенной по принципу организации и функционирования биологических нейронных сетей. Основными трудностями при их использовании являются необходимость большого количества обучающих данных, ресурсов и времени, а также их настройка: определение количества скрытых слоев, функции активации для каждого узла и пороговой ошибки. Однако нейронные сети способны отбирать признаки без участия человека, определять сложные зависимости между входными и выходными данными, а также адаптироваться к изменениям, благодаря чему нейронные сети зарекомендовали себя во многих областях, в том числе в анализе тональности текстов.

Наиболее распространенными нейронными сетями в области анализа тональности текстов являются *сверточные* и *рекуррентные*. Сверточные нейронные сети используют операцию свертки (текст разбивается на фрагменты, каждый из которых умножается на матрицу свертки поэлементно, после чего результат суммируется), которая происходит на соседних словах, что позволяет учесть контекст, например, отрицания. Рекуррентные нейронные сети имеют обратные связи от более удаленных элементов к менее, что позволяет сети учитывать не только текущие, но и предыдущие входные данные, благодаря чему вес каждого слова влияет на веса остальных слов в предложении.

2.3 Гибридные методы

Для объединения преимуществ лингвистического похода и методов машинного обучения разрабатываются *гибридные методы*, сочетающие в себе два и более из описанных методов [9]. Так, для решения задачи анализа тональности были объединены тональные словари и метод опорных векторов, сверточные нейронные сети и k-ближайших соседей [6], также в одной из работ была разработана гибридная система анализа тональности, объединившая методы на основе правил и сверточную нейронную сеть.

Сочетание сильных сторон подходов на основе лингвистики и машинного обучения позволяет получить более точные результаты, однако в то же время гибридные методы получают проблемы и ограничения каждого из подходов [12].

3 Классификация существующих решений

В данном разделе приводится иерархия существующих решений, предлагаются критерии оценки методов и проводится сравнение по выделенным критериям.

3.1 Иерархия методов

На основе приведенных в предыдущем разделе описаний можно составить иерархию методов анализа тональности естественно-языковых текстов на основе применяемых подходов. Данная иерархия представлена на рисунке 3.1.

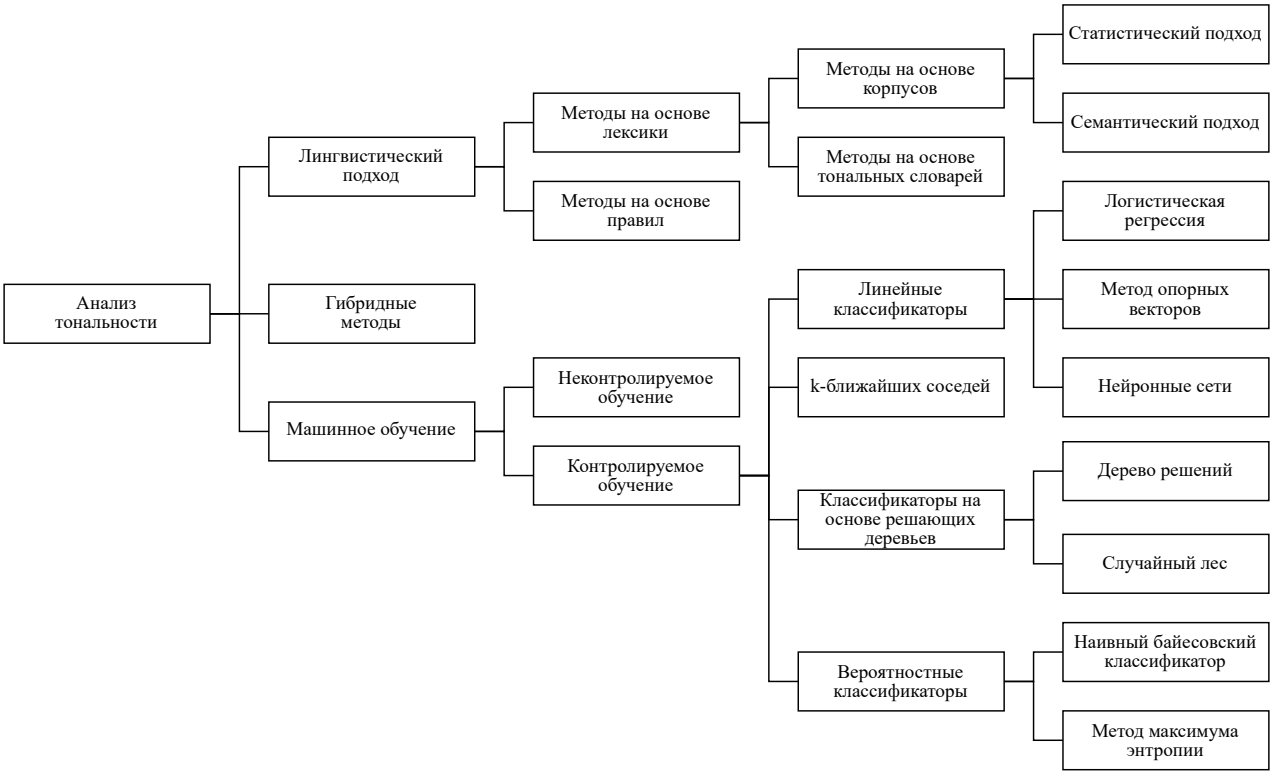

Рисунок 3.1 – Иерархия методов анализа тональности текстов

3.2 Сравнение и оценка методов

Анализ тональности текстов является задачей классификации, для которой традиционно используют такие метрики эффективности, как точность (precision), полнота (recall) и F-мера. Вычисление данных характеристик происходит на основе таблицы сопряженности (таблица 3.1), которая составляется

путем сравнения решения системы относительно класса текста с решением экспертов, которые формируют тестовые данные [18]. Далее приведенные выше величины называются метриками точности.

Таблица 3.1 – Таблица сопряженности

		Оценка эксперта	
		Положительная	Отрицательная
Оценка системы	Положительная	TP	FP
	Отрицательная	FN	TN

В таблице 3.1 используются следующие обозначения:

– TP (True Positive) — количество текстов, которые являются положительными и которые система определила, как положительные;

– FP (False Positive) — количество текстов, которые являются отрицательными и которые система определила, как положительные;

– TN (True Negative) — количество текстов, которые являются отрицательными и которые система определила, как отрицательные;

– FN (False Negative) — количество текстов, которые являются отрицательными и которые система определила, как положительные.

Эти же обозначения используются далее в формулах при определении метрик эффективности.

Точность (precision) — доля текстов, которые действительно принадлежат данному классу, относительно текстов, которые классификатор причислил к данному классу. Вычисляется по формуле (3.1).

$$precision = \frac{TP}{TP + FP} \tag{3.1}$$

Полнота (recall) — доля текстов, причисленных классификатором к данному классу, относительно всех текстов, принадлежащий ему в тестовой выборке. Вычисляется по формуле (3.2).

$$recall = \frac{TP}{TP + FN} \tag{3.2}$$

F-мера — среднее гармоническое точности и полноты, вычисляющееся по формуле (3.3).

$$F = \frac{2 \cdot precision \cdot recall}{recision + recall}, \tag{3.3}$$

где F — F-мера.

Также методы анализа тональности основываются на заранее подготовленных данных (набора слов, правил, коэффициентов и т. п.), с помощью которых происходит классификация. Такая подготовка требует дополнительных затрат и осуществляется различными путями [21], поэтому в качестве одного из критериев выбран способ проведения подготовки данных:

– ручной (данные подготавливаются людьми);

– автоматический (подготовка осуществляется с помощью вычислительной техники);

– смешанный (сначала осуществляется ручная подготовка данных, на основе которой проводится дополнительная автоматическая подготовка).

Требуемая во всех классификаторах предварительная разметка данных не учитывается.

При анализе тональности также может возникуть необходимость расширения предметной области текстов, классификация которых происходит, для чего может постребоваться повторная настройка или обучение классификатора [21]. Необходимость такой настройки в данной работе рассматривается как критерий оценки методов.

Результаты сравнения методов анилиза тональности приведены в таблице 3.2. Для краткости записи в данной таблице используются следующие обозначения описанных критериев:

– К1 — точность;

– К2 — полнота;

– К3 — F-мера;

– К4 — способ проведения подготовки данных;

– К5 — необходимость повторной настройки классификатора при расширении предметной области анализируемых текстов.

Таблица 3.2 – Сравнение методов анализа тональности

Метод	K1	K2	K3	K4	K5
На основе правил [10]	74.2%	73.7%	73.9%	ручной	требуется
На основе словарей [22]	97.6%	88.9%	93.1%	смешанный	требуется
На основе корпусов [23]	43.1%	94.3%	59.2%	смешанный	требуется
Наивный Байес [21]	75.5%	52.4%	67.9%	автоматический	не требуется
Логистическая регрессия [21]	76.9%	80.1%	78.5%	автоматический	не требуется
Максимум энтропии [19]	65.2%	98.1%	78.3%	автоматический	требуется
k-ближайших соседей [21]	59.1%	96.1%	73.2%	автоматический	не требуется
Дерево решений [21]	96.7%	64.5%	77.4%	автоматический	требуется
Случайный лес [21]	89.5%	76.5%	82.5%	автоматический	не требуется
Опорные векторы [19]	83.8%	83.9%	83.8%	автоматический	не требуется
Нейронные сети [21]	85.3%	88.4%	86.8%	смешанный	не требуется

3.3 Вывод

Таким образом, по данным таблицы 3.2 наиболее точным методом является метод на основе словарей, однако в нем подготовка данных осуществляется с участием человека, а расширение предметных областей текстов требует повторной настройки классификатора, как и в случае других методов лингвистического подхода, показывающих метрики точности, близкие к метрикам методов машинного обучения. Наиболее точным методом с автоматической

подготовкой данных для классификации и с отсутствием необходимости повторной настройки является метод опорных векторов. Для повышения точности классификации можно использовать нейронные сети, однако придется пожертвовать временем на их настройку. Остальные методы показывают приемлемые метрики точности в районе 60-80 процентов, что говорит об их применимости к решению конкретных задач для определенных предметных областей текстов.

ЗАКЛЮЧЕНИЕ

В ходе данной работы:

– были рассмотрены основные подходы к анализу тональности;

– были описаны методы аналаза тональности естественно-языковых текстов;

– были выделены критерии оценки, по которым проведено сравнение описанных методов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Самигулин Т. Р., Джурабаев А. Э. У. Анализ тональности текста методами машинного обучения // Научный результат. Информационные технологии. — 2021. — № 1. — URL: https://cyberleninka.ru/article/n/analiz-tonalnosti-teksta-metodami-mashinnogo-obucheniya (дата обращения: 15.12.2021).

2. Богданов А. Л., Дуля И. С. Сентимент-анализ коротких русскоязычных текстов в социальных медиа // Вестник Томского государственного университета. Экономика. — 2019. — № 47. — С. 220–241. — URL: https://cyberleninka.ru/article/n/sentiment-analiz-korotkih-russkoyazychnyh-tekstov-v-sotsialnyh-media (дата обращения: 15.12.2021).

3. Майорова Е. В. О сентимент-анализе и перспективах его применения // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 6, Языкознание: Реферативный журнал. — 2020. — № 4. — С. 78–87. — URL: https://cyberleninka.ru/article/n/o-sentiment-analize-i-perspektivah-ego-primeneniya (дата обращения: 15.12.2021).

4. Sharma A. Natural Language Processing and Sentiment Analysis // International Research Journal of Computer Science. — 2021. — Vol. 8. — P. 237–242. — URL: https://www.researchgate.net/publication/355927843_NATURAL_LANGUAGE_PROCESSING_AND_SENTIMENT_ANALYSIS (дата обращения: 15.12.2021).

5. Колмогорова А. В. Использование текстов жанра «Интернет-откровение» в контексте решения задач сентимент-анализа // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. — 2019. — № 3. — С. 71–82. — URL: https://cyberleninka.ru/article/n/ispolzovanie-tekstov-zhanra-internet-otkrovenie-v-kontekste-resheniya-zadach-sentiment-analiza (дата обращения: 15.12.2021).

6. Двойникова А. А., Карпов А. А. Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных // Информационно-управляющие системы. — 2020. — С. 20–30. — URL: https://cyberleninka.ru/article/n/analiticheskiy-obzor-podhodov-k-raspoznavaniyu-tonalnosti-russkoyazychnyh-tekstovyh-dannyh (дата обращения: 16.12.2021).

7. Pradha S., Halgamuge M. N., Vinh N. T. Q. Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data // 2019 11th International Conference on Knowledge and Systems Engineering (KSE). — 2019. — P. 1–8. — URL: https://www.semanticscholar.org/paper/Effective-Text-Data-Preprocessing-Technique-for-in-Pradha-Halgamuge/2efa3f13d09ac7954bddd4b7a190c47d144c533f (дата обращения: 16.12.2021).

8. Sentiment analysis using logistic regression algorithm / Y. Jaswanth, R. M. S. Kumar, R. M. Sudhan [et al.] // European Journal of Molecular & Clinical Medicine. — 2020. — Vol. 7. — URL: https://www.semanticscholar.org/paper/Sentiment-analysis-using-logistic-regression-Jaswanth-Kumar/1af4aaa6670a8bf62460ef69476ead4f984993af (дата обращения: 15.12.2021).

9. Sentiment Analysis for Mining Texts and Social Networks Data: Methods and Tools / C. Zucco, B. Calabrese, G. Agapito [et al.] // WIREs Data Mining and Knowledge Discovery. — 2020. — Vol. 10, no. 1. — URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1333 (дата обращения: 15.12.2021).

10. Khalil E., El Houby E., Mohamed H. Sentiment Analysis Tasks and Approaches // International Journal of Computer Science and Information Security. — 2021. — Vol. 19. — URL: https://www.researchgate.net/publication/356068600_Sentiment_Analysis_Tasks_and_Approaches (дата обращения: 15.12.2021).

11. Sentiment Analysis Techniques for Social Media Data: A Review / D. Sharma,

M. Sabharwal, V. Goyal [et al.] // First International Conference on Sustainable Technologies for Computational Intelligence. — 2020. — P. 75–90. — URL: https://www.researchgate.net/publication/336988754_Sentiment_Analysis_Techniques_for_Social_Media_Data_A_Review (дата обращения: 16.12.2021).

12. Smetanin S. The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives // IEEE Access. — 2020. — Vol. 08. — P. 110693–1110719. — URL: https://www.researchgate.net/publication/342193894_The_Applications_of_Sentiment_Analysis_for_Russian_Language_Texts_Current_Challenges_and_Future_Perspectives (дата обращения: 16.12.2021).

13. Berka P. Sentiment analysis using rulebased and casebased reasoning // Journal of Intelligent Information Systems. — 2020. — Vol. 55. — P. 51–66. — URL: https://link.springer.com/article/10.1007/s10844-019-00591-8#citeas (дата обращения: 16.12.2021).

14. Pathak A., Sharma S., Pandey R. A Methodological Survey on Sentiment Analysis Techniques and Their Applications in Opinion Mining // International Journal of Emerging Trends in Engineering and Development. — 2021. — Vol. 1. — P. 37–45. — URL: https://www.researchgate.net/publication/349154400_A_METHODOLOGICAL_SURVEY_ON_SENTIMENT_ANALYSIS_TECHNIQUES_AND_THEIR_APPLICATIONS_IN_OPINION_MINING (дата обращения: 15.12.2021).

15. Sentiment analysis of online product reviews using Lexical Semantic Corpus-Based technique / R. Aminuddin, A. Z. Zulkefli, N. A. Moketar [et al.] // 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE). — 2021. — P. 233–238. — URL: https://www.semanticscholar.org/paper/Sentiment-analysis-of-online-product-reviews-using-Aminuddin-Zulkefli/0ca4c7449a42b64f16aaf4bdce433a2b16953a2e (дата обращения: 16.12.2021).

16. Mehta P., Pandya D. A Review On Sentiment Analysis Methodologies, Practices And Applications // International Journal of Scientific & Technology Research. — 2020. — Vol. 9. — P. 601–609. — URL: https://www.researchgate.net/publication/344487215_A_Review_On_Sentiment_Analysis_Methodologies_Practices_And_Applications (дата обращения: 16.12.2021).

17. Berrar D. Bayes' Theorem and Naive Bayes Classifier. — 2018. — URL: https://www.researchgate.net/publication/324933572_Bayes%27_Theorem_and_Naive_Bayes_Classifier (дата обращения: 15.12.2021).

18. Alvi N., Talukder K. H. Sentiment Analysis of Bengali Text using CountVectorizer with Logistic Regression // 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). — 2021. — P. 01–05. — URL: https://www.semanticscholar.org/paper/Sentiment-Analysis-of-Bengali-Text-using-with-Alvi-Talukder/fecffbf1db85b91fc598fd395817685fff6b3df1 (дата обращения: 16.12.2021).

19. Gritta M. A Comparison of Techniques for Sentiment Classification of Film Reviews // CoRR. — 2019. — URL: http://arxiv.org/abs/1905.04727 (дата обращения: 16.12.2021).

20. N-Gram and K-Nearest Neighbor Algorithm for Sentiment Analysis on Capital Relocation / M. Ramadhon, A. Arini, F. Mintarsih [et al.] // International Conference on Cyber and IT Service Management (CITSM). — 2021. 09. — P. 1–6. — URL: https://www.researchgate.net/publication/356149710_N-Gram_and_K-Nearest_Neighbor_Algorithm_for_Sentiment_Analysis_on_Capital_Relocation (дата обращения: 16.12.2021).

21. Батура Т. В. Методы автоматической классификации текстов // Международный журнал Программные продукты и системы. — 2017. 03. — Т. 23. — С. 85–99. — URL: https://www.researchgate.net/publication/315328102_Metody_avtomaticeskoj_klassifikacii_tekstov (дата обращения: 16.12.2021).

22. Wilis K., Hidayatullah H., Parasian S. The Accuracy Comparison of Social Media Sentiment Analysis Using Lexicon Based and Support Vector Machine on Souvenir Recommendations // TEST Engineering & Management. — 2020. — Vol. 83. — P. 3953–3961. — URL: http://eprints.upnyk.ac.id/23088/1/3.theaccuracycomparison.pdf (дата обращения: 16.12.2021).

23. Almatarneh Sattam, Gamallo Pablo. A lexicon based method to search for extreme opinions // PLoS ONE. — 2018. 05. — T. 13. — URL: https://www.researchgate.net/publication/325370564_A_lexicon_based_method_to_search_for_extreme_opinions (дата обращения: 16.12.2021).