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

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

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

# 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 и обосновываются 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 использование носителями средств выразительности

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):

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

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

$\theta$ — вектор-столбец коэффициентов, получаемых в ходе обработки обучающей выборки.

Для определения класса текст представляется в виде вектора-столбца $x$, и 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].

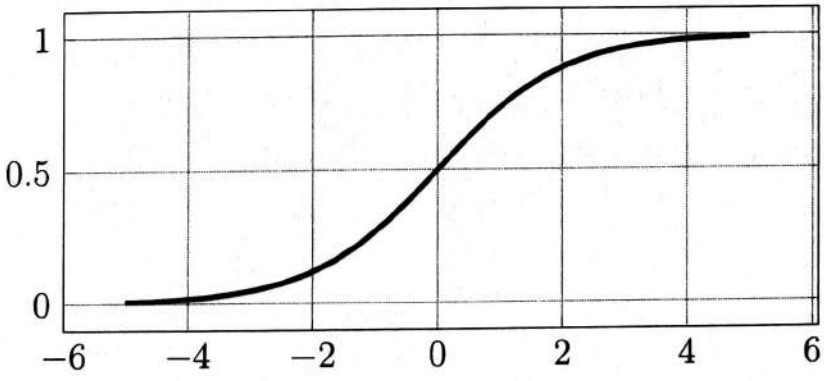


Рисунок 2.1 – График логистической функции

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. В качестве расстояния используют косинусное сходство (формула (2.11)), соответствующее косинусу угла $\theta$ между векторами $\vec{A}$ и $\vec{B}$, которые задают сравниваемые тексты.

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

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-

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).

ской подготовкой данных для классификации и с отсутствием необходимости повторной настройки являются нейронные сети и метод опорных векторов. 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.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1. Самигулин Т. Р., Джурабаев А. Э. У. Анализ тональности текста методами машинного обучения // Научный результат. Информационные технологии. — Белгород, 2021. — № 1. — С. 55–62. — URL: https://cyberleninka.ru/article/n/analiz-tonalnosti-teksta-metodami-mashinnogo-obucheniya.

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

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

4. Sharma A. Natural Language Processing and Sentiment Analysis // International Research Journal of Computer Science. — 2021. — Vol. 8. — P. 237–242. — URL: http://www.irjcs.com/volumes/Vol8/iss-10/01.OCCS10080.pdf.

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

6. Двойникова А. А., Карпов А. А. Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных // Информационно-управляющие системы. — СПб., 2020. — № 4. — С. 20–30. — URL: https://cyberleninka.ru/article/