



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:	<u>ИУ7-53Б</u> (группа)	_____ (подпись, дата)	<u>М. Д. Маслова</u> (И. О. Фамилия)
Руководитель:		_____ (подпись, дата)	<u>А. А. Оленев</u> (И. О. Фамилия)

2022 г.

## **РЕФЕРАТ**

Расчетно-пояснительная записка 14 с., 0 рис., 0 табл., 12 источн., 4 прил.  
**АНАЛИЗ ТОНАЛЬНОСТИ**

# СОДЕРЖАНИЕ

<b>РЕФЕРАТ</b>	<b>2</b>
<b>ВВЕДЕНИЕ</b>	<b>4</b>
<b>1 Анализ предметной области</b>	<b>5</b>
1.1 Актуальность задачи . . . . .	5
1.2 Основные определения . . . . .	6
1.3 Формализация задачи . . . . .	7
<b>2 Описание существующих решений</b>	<b>8</b>
2.1 Лингвистический подход . . . . .	8
2.1.1 Методы, основанные на правилах . . . . .	8
2.1.2 Методы, основанные на тональных словарях . . . . .	9
2.1.3 Методы, основанные на корпусах . . . . .	9
2.2 Методы машинного обучения . . . . .	9
2.2.1 Наивный Байес . . . . .	9
2.2.2 Логическая регрессия . . . . .	9
2.2.3 k-ближайших соседей . . . . .	9
2.2.4 что-то там про лес и деревья ;) . . . . .	9
2.2.5 Нейронные сети . . . . .	9
2.3 Гибридные . . . . .	9
<b>3 Классификация существующих решений</b>	<b>10</b>
3.1 Технология метода . . . . .	10
3.2 Уровни . . . . .	10
3.3 Скорость . . . . .	10
3.4 Данные/память . . . . .	10
3.5 Точность . . . . .	10
3.6 Время разработки . . . . .	10
3.7 По предварительной обработке данных . . . . .	10
<b>4 Заключение</b>	<b>11</b>
<b>СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ</b>	<b>14</b>

# **ВВЕДЕНИЕ**

# **1 Анализ предметной области**

## **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 виде текстовой информации, поэтому становится актуальной задача анализа текстов на естественном языке. [1] Одной из этих задач является анализ тональности и 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 [2] 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, принятых для борьбы с вирусом COVID-19, на жизнь людей) [3];
- исследование рынка и прогнозирование цен на акции [3];
- выявление случаев эмоционального насилия и пресечение противоправных действий [4].

Решение описанных задач требует анализ большого количества текстов, что делает невозможным их ручную обработку. Также при оценке тональности текста человеком трудно соблюсти критерии этой оценки. 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 и переносных значений слов и фраз. Также основной из проблем сентимент-анализа является разная окраска одного и того слова в текстах на различные тематики: слово, которое считается положительным в одной, в то же время считается отрицательным в другой.

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

**ССЫЛКИ!!!**

## **1.2 Основные определения**

**Анализ тональности текста** (sentiment analysis) – 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 [5].

В настоящее время выделяют три основных подхода к определению тональности текста [5]:

- *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** к решению задачи анализа тональности, во всех **подходах** требуется предварительная обработка текста, основными этапами которой являются:

- приведение текста к *единому регистру* для сокращения количества слов, которые необходимо хранить одновременно [6];
- *удаление пунктуации и шума* (упоминаний пользователей, ссылок, хештегов) [6];
- *токенизация* или разбиение исходного текста на лексемы, в простейшем случае – разбиение по пробельным символам [7];
- *удаление стоп-слов*, то есть слов не несущих никакой смысловой нагрузки, с целью повышения точности [6];
- *стемминг* или *лемматизация* – приемы приведения слов форм слова к общему виду; в случае стеммига 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 [6];

- обработка отрицаний [8].

### 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**

### **2.1 Лингвистический подход**

Методы использующие лингвистический подход можно разделить на три основные категории:

- методы на основе правил;
- методы на основе словарей;
- методы на основе корпусов.

#### **2.1.1 Методы, основанные на правилах**

Работа **методов на основе правил** реализуется с помощью большого набора созданных в ручную правил конструкции "если  $\rightarrow$  то"[9].

Данные алгоритмы имеют отличную производительность 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 [10].

В целях ускорения процесса разработки для создания набора правил может использоваться машинное обучение, поэтому в некоторых научных работах [11] [12] данные методы относят к методам машинного обучения.



### **2.1.2 Методы, основанные на тональных словарях**

### **2.1.3 Методы, основанные на корпусах**

## **2.2 Методы машинного обучения**

### **2.2.1 Наивный Байес**

### **2.2.2 Логическая регрессия**

### **2.2.3 k-ближайших соседей**

### **2.2.4 что-то там про лес и деревья ;)**

### **2.2.5 Нейронные сети**

## **2.3 Гибридные**

Общее описание

### **3 Классификация существующих решений**

#### **3.1 Технология метода**

machine learning

lexicon approach

hybrid

#### **3.2 Уровни**

Здесь необходимо пояснить за семантические связи. "Еда вкусная, но обслуживание так себе". В целом – скорее всего нейтральный, но по аспектам: о еде: положительно, обобслуживании: отрицательно.

document

sentence

approach

#### **3.3 Скорость**

#### **3.4 Данные/память**

#### **3.5 Точность**

#### **3.6 Время разработки**

#### **3.7 По предварительной обработке данных**

## **4 Заключение**

В ходе данной работы было выявлено:

- преобладание методов машинного обучения в данной сфере за счет ...;

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Богданов А. Л., Дуля И. С.* Сентимент-анализ коротких русскоязычных текстов в социальных медиа // Вестн. Том. гос. ун-та. 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).
2. *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).
3. *Sharma A.* Natural Language Processing and Sentiment Analysis // International Research Journal of Computer Science. — 2021. — Т. 8. — С. 237–242. — URL: [https://www.researchgate.net/publication/355927843\\_NATURAL\\_LANGUAGE\\_PROCESSING\\_AND\\_SENTIMENT\\_ANALYSIS](https://www.researchgate.net/publication/355927843_NATURAL_LANGUAGE_PROCESSING_AND_SENTIMENT_ANALYSIS) (дата обращения: 15.12.2021).
4. *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).
5. *Самигулин Т. Р., Джурбаев 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).
6. *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. — С. 1–8. — URL: <https://www.>

semanticscholar . org / paper / Effective - Text - Data - Preprocessing - Technique - for - in - Pradha - Halgamuge / 2efa3f13d09ac7954bddd4b7a190c47d144c533f (дата обращения: 16.12.2021).

7. Sentiment analysis using logistic regression algorithm / Y. Jaswanth [и др.] // Т. 7. — 2020. — URL: [https : / / www . semanticscholar . org / paper / Sentiment - analysis - using - logistic - regression - Jaswanth - Kumar / 1af4aaa6670a8bf62460ef69476ead4f984993af](https://www.semanticscholar.org/paper/Sentiment-analysis-using-logistic-regression-Jaswanth-Kumar/1af4aaa6670a8bf62460ef69476ead4f984993af) (дата обращения: 15.12.2021).
8. Sentiment analysis for mining texts and social networks data: Methods and tools / C. Zucco [и др.] // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. — 2020. — Т. 10. — URL: [https : / / www . semanticscholar . org / paper / Sentiment - analysis - for - mining - texts - and - social - and - Zucco - Calabrese / 8e3f93b6dd166db7843c4c8cbc2393a8e177d455](https://www.semanticscholar.org/paper/Sentiment-analysis-for-mining-texts-and-social-and-Zucco-Calabrese/8e3f93b6dd166db7843c4c8cbc2393a8e177d455) (дата обращения: 16.12.2021).
9. 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](https://cyberleninka.ru/article/n/analiticheskiy-obzor-podhodov-k-raspoznavaniyu-tonalnosti-russkoyazychnyh-tekstovyh-dannyh) (дата обращения: 16.12.2021).
10. Smetanin S. The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives // IEEE Access. — 2020. — Т. 08. — С. 110693—1110719. — URL: [https : / / www . researchgate . net / publication / 342193894 \\_ The \\_ Applications \\_ of \\_ Sentiment \\_ Analysis \\_ for \\_ Russian \\_ Language \\_ Texts \\_ Current \\_ Challenges \\_ and \\_ Future \\_ Perspectives](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).
11. Sentiment Analysis Techniques for Social Media Data: A Review / D. Sharma [и др.] // First International Conference on Sustainable Technologies for

Computational Intelligence. — 2020. — С. 75—90. — URL: [https://www.researchgate.net/publication/336988754\\_Sentiment\\_Analysis\\_Techniques\\_for\\_Social\\_Media\\_Data\\_A\\_Review](https://www.researchgate.net/publication/336988754_Sentiment_Analysis_Techniques_for_Social_Media_Data_A_Review) (дата обращения: 16.12.2021).

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