



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 10 с., 0 рис., 0 табл., 5 источн., 4 прил.
АНАЛИЗ ТОНАЛЬНОСТИ

СОДЕРЖАНИЕ

РЕФЕРАТ	2
ВВЕДЕНИЕ	4
1 Анализ предметной области	5
1.1 Основные определения	5
1.2 Актуальность задачи	5
1.3 Формализация задачи	6
1.4 Препроцессинг	6
2 Описание существующих решений	7
2.1 Методы основанные на лексике	7
2.2 Методы машинного обучения	7
2.2.1 Наивный Байес	7
2.2.2 Логическая регрессия	7
2.2.3 к ближайших соседей	7
2.2.4 что-то там про лес и деревья ;)	7
2.2.5 Нейронки	7
2.3 Гибридные	7
3 Классификация существующих решений	8
3.1 Технология метода	8
3.2 Уровни	8
3.3 От тематики???	8
3.4 Скорость	8
3.5 Данные/память	8
3.6 Точность	8
3.7 Время разработки	8
3.8 Язык???	8
3.9 По предварительной обработке данных	8
3.10 Влияние частей речи???	8
4 Заключение	9

ВВЕДЕНИЕ

1 Анализ предметной области

1.1 Основные определения

Sentiment analysis (SA) или анализ мнений извлекает и анализирует субъективную информацию из различных источников, таких как Интернет, социальные сети и другие источники, чтобы определить мнение людей с помощью обработки естественного языка (NLP), вычислительной лингвистики и анализа текста. Эта проанализированная информация дает представление о чувствах или отношении общества к определенным предметам, лицам или идеям и определяет контекстуальную полярность информации. [1]

Тональность

Анализ тональности

Предобработка

Корпус

Словарь

Позитивный/негативный/нейтральный

1.2 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. [2] 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 [3] 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 [4];
- анализ политических и социальных взглядов пользователей (например, влияние мер, предпринятых для борьбы с вирусом COVID-19, на жизнь людей) [4];

- исследование рынка и прогнозирование цен на акции [4];
- выявление случаев эмоционального насилия и пресечение противоправных действий [5].

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 использования носителями средств выразительности и переносных значений слов и фраз. Также основной из проблем sentiment-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.

1.3 Формализация задачи

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

- положительный;
- отрицательный;
- нейтральный.

При этом определяется лишь **факт?** принадлежности тому или иному классу без определения вероятности "чего-то там"

1.4 Препроцессинг

единый регистр

удаление пунктуации

лемматизация/стемминг

удаление стоп слов

удаление шума (хэш-тэгов, ссылок и тд)

2 Описание существующих решений

2.1 Методы основанные на лексике

Надо найти

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 Нейронки

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 От тематики???

3.4 Скорость

3.5 Данные/память

3.6 Точность

3.7 Время разработки

3.8 Язык???

3.9 По предварительной обработке данных

3.10 Влияние частей речи???

4 Заключение

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Khalil E., El Houby E., Mohamed H.* Sentiment Analysis Tasks and Approaches // International Journal of Computer Science and Information Security. — 2021. — Т. 19. — URL: https://www.researchgate.net/publication/356068600_Sentiment_Analysis_Tasks_and_Approaches (дата обращения: 15.12.2021).
2. *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).
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. — Т. 8. — С. 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).