

# Анализ тональности отзывов с помощью методов машинного обучения

Верхотуров М.А.

Профессор, д.т.н.

Институт информатики, математики и  
робототехники

Уфимский университет науки и технологий  
Уфа, Россия

e-mail: verhotur\_m@rambler.ru

Верхотурова Г.Н.

Доцент, к.т.н.

Институт информатики, математики и  
робототехники

Уфимский университет науки и технологий  
Уфа, Россия

e-mail: verhoturova.gn@yandex.ru

Чельшкина Е.М.

Уфимский университет науки и технологий

Уфа, Россия

e-mail: chelyshkinaekaterina@gmail.com

## Аннотация<sup>1</sup>

В работе рассматривается проблема проведения анализа тональности текста (сентимент-анализа) при оценке отзывов клиентов в таких областях, как экономика, менеджмент, маркетинг, политология. Особенностью задачи является то, что точность оценки тональности напрямую зависит от конкретного языка и его особенностей. Для решения данной проблемы использовался подход с использованием искусственных нейронных сетей. В статье описана разработанная модель нейронной сети, ее архитектура и гиперпараметры, используемые инструментальные средства, примеры работы программного обеспечения, а также результаты вычислительного эксперимента.

**Ключевые слова:** тональность текста; сентимент-анализ; рекуррентная нейронная сеть; токенизация; лемматизация.

## 1. Введение

Современные компании обладают большими объемами текстовых данных, такими как отзывы, электронные письма, комментарии в социальных сетях, упоминания в средствах массовой информации и электронные письма, содержание которых чрезвычайно важно при анализе и изучении удовлетворенности клиентов бренда. В отчете компании NICE, которая специализируется на программном обеспечении для управления взаимоотношениями с клиентами, отмечается, что

больше 50% компаний считают, что их клиенты довольны обслуживанием и положительно относятся к их бренду, однако только 15% клиентов согласны с этим. Отчет демонстрирует значительные расхождения между восприятием клиентского опыта со стороны самих клиентов и тем, насколько серьезно бизнес-руководители воспринимают их ожидания. Результаты также показывают, что эти разрывы оказывают серьезное влияние на лояльность к бренду – 57% клиентов отказываются от бренда всего после одного или двух негативных взаимодействий с ним [1]. Кроме того, по данным исследований, инструменты случайных рекомендаций обеспечивают удовлетворенность лишь 65% клиентов. В то время как предложения, основанные на предыдущем опыте и впечатлениях клиентов, помогают повысить этот показатель до 90% [2].

## 2. Анализ тональности текста

Анализ тональности текста или сентимент-анализ – процесс определения эмоциональной окраски текста, которая выражается в позитивном, негативном, и иногда нейтральном, отношении к чему-либо. Под тональностью обычно подразумевается эмоциональная составляющая мнения автора, которая представляет собой полярность оценки [3]. Часто выделяют три категории тональности:

- Положительная – выражение положительных эмоций, таких как одобрение, удовлетворение, радость.
- Негативная – выражение негативных эмоций, таких как недовольство, грусть, гнев.
- Нейтральная – отсутствие какого-либо тонального компонента в высказывании, иногда выделяется как промежуточная тональность между положительной и негативной.

---

Труды X Международной научной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 12-14 ноября, Уфа-Баку-Чандигарх, 2024

Анализ тональности текста применяют в таких областях, как экономика, менеджмент, маркетинг, политология и социальные науки для решения следующих задач:

- Контроль репутации бренда. Анализ тональности применяется для оценки реакции потребителей на товары, услуги и маркетинговые кампании для улучшения имиджа и качества товаров и услуг.
- Прогнозирование клиентских предпочтений и рыночных тенденций. Путем анализа тональности текстов компании могут предсказать предпочтения и потребности клиентов, адаптируя под их запросы свои товары и услуги.
- Контроль социально-политической обстановки. С помощью анализа тональности текстов можно воспроизводить мониторинг общественного мнения и настроения.
- Фильтрация контента. Анализ тональности текста может применяться для автоматической фильтрации потенциально негативной или вредоносной информации.

Поскольку компьютерная лингвистика работает с естественными языками, существуют некоторые проблемы, которые возникают при разработке и применении алгоритмов для определения тональности контента:

– высокая синтаксическая сложность языка – различные сложные грамматические конструкции, являющиеся особенностью языка, могут повлиять на точность анализа тональности текста [4];

– наличие в тексте сарказма, иронии и двусмысленности – одна из больших проблем анализа тональности текста, поскольку слова в данном контексте используются в переносном, а не в прямом, значении, что значительно затрудняет процесс анализа.

## 2.1. Методы решения задачи анализа тональности текста

На **Ошибка! Источник ссылки не найден.**1 представлены подходы для решения задачи анализа тональности текста [5].

Для решения задачи анализа тональности текста часто применяется подход обучения с учителем. Обучающие наборы данных делятся на классы: позитивный и негативный, – и затем используются для обучения классификатора. Наиболее распространенными методами в этом случае являются метод опорных векторов (SVM) и наивный байесовский классификатор. Обучение без учителя в этой области применяется редко,

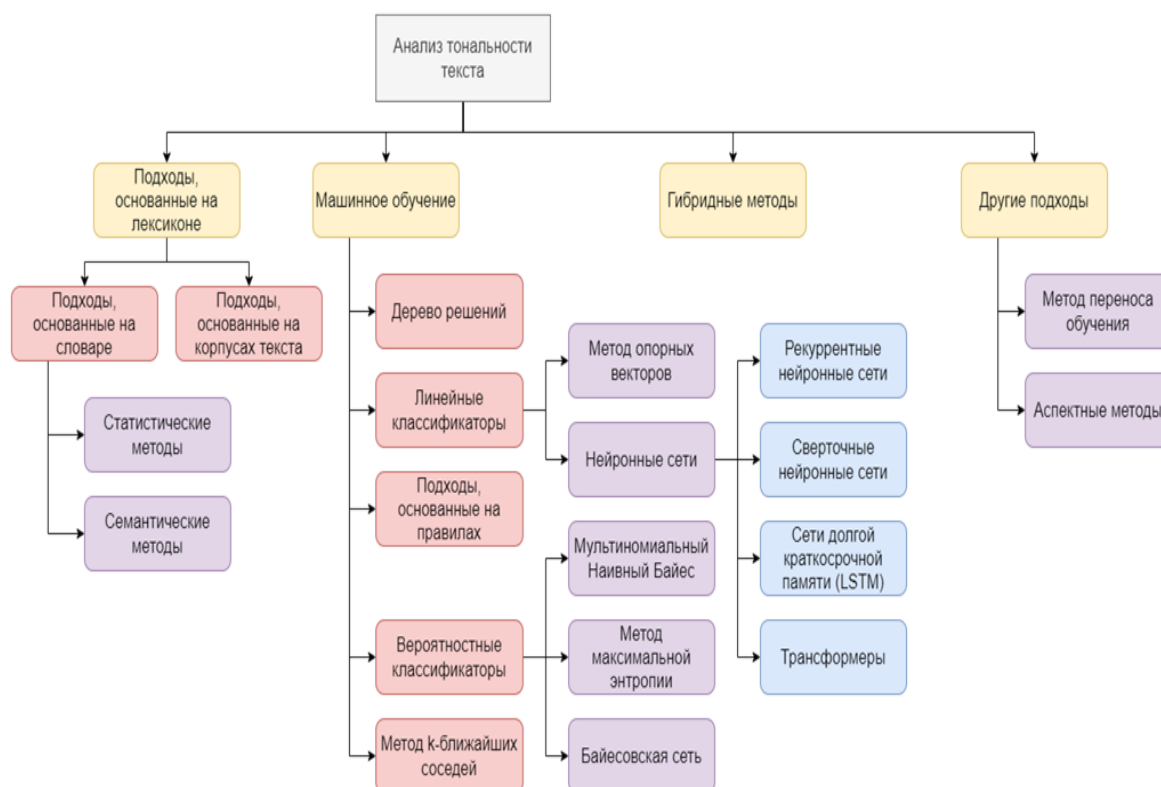


Рис. 1. Подходы и методы решения задачи анализа тональности текста

поскольку подразумевает автоматическую кластеризацию данных, что значительно снижает точность определения тональности по сравнению с

методами обучения с учителем, где данные заранее разделяют на классы [4].

Среди методов машинного обучения наилучшей точности при классификации текста по тональности удалось добиться с помощью искусственных нейронных сетей, показатель точности составил более 85% [6].

Точность оценки тональности напрямую зависит от конкретного языка и его особенностей. На данный момент системы для анализа русскоязычных текстов развиты меньше, чем для некоторых других мировых языков и результаты sentiment-анализа русскоязычных текстов ниже, чем англоязычных. Это связано с тем, что русский язык имеет достаточно сложную структуру. Поэтому актуальной проблемой является разработка алгоритмов для повышения точности определения эмоциональной окраски, учитывающие специфику и особенности языка [7,8].

## 2.2. Применение нейронных сетей при анализе тональности текста

Анализ тональности текста является одной из задач обработки естественного языка, поэтому контекст и порядок слов крайне важны при анализе. Сеть долгой краткосрочной памяти (LSTM) – рекуррентная нейронная сеть, которая способна сохранять информацию о уже обработанных словах для дальнейшего анализа, потому именно эта архитектура была выбрана для разработки [9,10]. LSTM-модуль подходит для работы с естественным языком, в том числе и для анализа тональности, благодаря своей способности моделировать долгосрочные зависимости и учитывать контекст и порядок слов в предложении. В процессе исследования была разработана модель нейронной сети, которая состоит из следующих слоев:

- входной слой;
- embedding-слой, который представляет токены в виде векторов;
- LSTM-слой;
- dropout-слой, который используется для предотвращения переобучения;
- линейный слой;
- sigmoid-слой;
- выходной слой.

Для компиляции архитектуры модели в данной работе были выбраны следующие гиперпараметры:

– функция потерь – бинарная кросс-энтропия, которая рассчитывается по формуле (1):

$$BCE(p, y) = -(y \log p + (1 - y) \log(1 - p)), (1)$$

где  $y$  – истинная метка образца (0 или 1);

$p$  – спрогнозированная моделью вероятность того, что образец относится к классу 1;

– метод оптимизации – алгоритм адаптивной оптимизации Adam с шагом 0,001;

– размер пакета (batch size) – 100 образцов, обрабатываемых перед обновлением весов модели;

– количество эпох (epochs) – 3 эпохи, которых достаточно для обучения модели без риска переобучения.

Функция потерь нейронной сети используется для оценки качества работы сети. Она вычисляет разницу между предсказанными значениями и реальными метками, которые модель должна предсказать. Цель обучения нейронной сети заключается в минимизации этой функции потерь. Функция потерь выбирается в соответствии с типом задачи, которую необходимо решить. Для задач классификации часто используется кросс-энтропия. Для бинарной классификации, когда выходные данные могут быть представлены как вероятность принадлежности к одному из двух классов, используется частный случай кросс-энтропии – бинарная кросс-энтропия.

Точность определения показывает, какой процент предсказанных положительных примеров на самом деле положителен.

Полнота отражает способность модели находить все положительные примеры в наборе данных. Это доля истинно положительных случаев от всех фактических положительных примеров.

F-мера представляет собой гармоническое среднее между точностью (precision) и полнотой (recall).

В качестве обучающей выборки были выбраны три набора данных, состоящих из отзывов на различные товары и услуги [11,12]:

– RuReviews: An Automatically Annotated Sentiment Analysis Dataset for Product Reviews in Russian – отзывы на женскую одежду и аксессуары;

– Russian Bank Reviews – отзывы клиентов различных банков;

– Russian Hotel Reviews Dataset – отзывы постояльцев отелей.

Из данных наборов были выбраны только положительные и отрицательные отзывы, нейтральные отзывы в обучении и анализе не участвовали.

Репрезентативная выборка содержала 40 тысяч положительных и 40 тысяч отрицательных отзывов, предварительно была перемешана и разделена на три части [13]:

– обучающая – 64 тысячи отзывов (80% всех отзывов);

– валидационная – 8 тысяч отзывов (10% всех отзывов);

– тестовая – 8 тысяч отзывов (10% всех отзывов).

Перед обучением нейронной сети или классификацией необходимо подготовить данные, так как нейронные сети умеют работать только с

числовыми векторами. Этап предварительной обработки текста является крайне важным и включает в себя следующие действия [14]:

- преобразование текста в нижний регистр;
- удаление знаков препинания;
- токенизация;
- удаление стоп-слов;
- лемматизация или стемминг.

Токенизация – процесс разбиения текста на список токенов, которые представляют собой слова, предложения или отдельные символы. В данной работе под токенами подразумеваются отдельные слова.

Удаление стоп-слов – процесс удаления общеупотребительных слов, которые несут мало смысловой нагрузки. В русском языке стоп-словами являются предлоги, союзы, местоимения, частицы и другие распространенные слова.

Стемминг – это процесс нахождения основы слова. При стемминге удаляются все аффиксы слова и остается только его основа, которая не всегда совпадает с корнем слова. Однако русский язык имеет сложную морфологическую изменяемость слов, которая может стать источником критических ошибок при использовании стемминга в алгоритме предварительной обработки текста. В качестве решения данной проблемы часто используют лемматизацию.

Лемматизация – это процесс приведения слова к его лемме – словарной (нормальной) форме. Лемматизация показывает исключительные результаты при использовании ее в тех задачах обработки естественного языка, где важно учитывать значение слов в контексте. Таким образом, при решении задач классификации текста по тональности в предварительной обработке предпочтительнее использовать лемматизацию для повышения точности классификации.

На первом этапе входные данные проходят процесс векторизации – преобразование слов в их векторное представление. Для векторизации использовалась модель Fasttex. Модель строит векторные представления n-грамм и складывает все векторные представления n-грамм, тем самым получая векторное представление слова [15].

На следующем этапе осуществляется анализ тональности отзывов с помощью обученной рекуррентной нейронной сети LSTM на основе входных данных [16]. Вероятность принадлежности к положительному классу рассчитывается в слое-сигмоиде по следующему принципу: если принадлежность к положительному классу меньше 0,5, то значит, что отзыв относится к отрицательному классу.

Программа реализована на языке Python версии 3.11 с использованием следующих дополнительных

библиотек: TensorFlow, Keras, Scikit-learn, PyTorch, Flask, PyMorphy3, Pandas, Chardet, Numpy.

### 3. Вычислительный эксперимент

Оценка качества классификации программного продукта проводилась на тестовой выборке, которая не участвовала в обучении модели [17]. Для оценки качества классификации рассчитывались следующие метрики:

- доля правильных ответов (accuracy),
- точность (precision),
- полнота (recall),
- F-мера.

Точность определения показывает, какой процент предсказанных положительных примеров на самом деле положителен. Полнота отражает способность модели находить все положительные примеры в наборе данных. Это доля истинно положительных случаев от всех фактических положительных примеров. F-мера представляет собой гармоническое среднее между точностью (precision) и полнотой (recall).

В таблице 1 представлены значения каждой метрики качества для обученной нейронной сети.

Таблица 1. – Метрики качества нейронной сети

| Метрика качества                   | Показатель |
|------------------------------------|------------|
| Точность (precision)               | 0,9005     |
| Полнота (recall)                   | 0,932      |
| Мера F1                            | 0,916      |
| Доля правильных ответов (accuracy) | 0,9123     |

Матрица ошибок для нейронной сети представлена на рисунке 2.

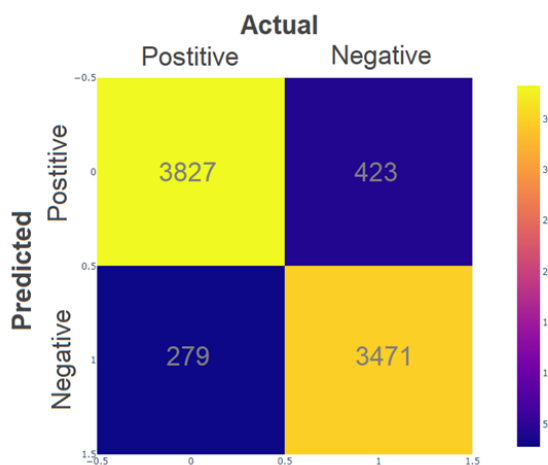


Рис. 2. Матрица ошибок

### 3. Заключение

Анализ результатов проведенных вычислительных экспериментов показал, что доля правильных ответов на валидационной выборке составила 91%. Этот показатель можно улучшить, например, путем дообучения модели распознавания эмоциональной окраски текста, в котором присутствует сарказм или ирония, на специально подготовленном наборе данных с отзывами, содержащими сарказм и иронию, или путем добавления собственного тонального словаря. Точность классификации также можно повысить, обучив нейронную сеть на таком наборе данных, который будет отражать отзывы на различные товары и впечатления потребителей других сфер обслуживания.

### Список используемых источников

1. Digital-First Customer Experience Report [Электронный ресурс] / NICE CXone. – 2022. – URL: <https://get.nice.com/Digital-CX-Research-Report.html> (дата обращения: 18.04.2024).
2. Rosa, R. L. Music recommendation system based on user's sentiments extracted from social networks / R. L. Rosa, D. Z. Rodriguez and G. Bressan. – IEEE Transactions on Consumer Electronics, 2015. – Vol. 61,3. – Pp. 359-367. – DOI: 10.1109/TCE.2015.7298296.
3. Семина Т.А. Анализ тональности текста: современные подходы и существующие проблемы [Электронный ресурс] / Семина Т.А. – Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 6, Языкознание: Реферативный журнал, 2020. – №4. – URL: <https://cyberleninka.ru/article/n/analiz-tonalnosti-teksta-sovremennye-podhody-i-suschestvuyuschie-problemy> (дата обращения: 15.04.2024).
4. Двойникова, А.А. Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных [Электронный ресурс] / Двойникова А.А., Карпов А.А. – Информационно-управляющие системы, 2020. – №4 (107). – URL: <https://cyberleninka.ru/article/n/analiticheskiy-obzor-podhodov-k-raspoznavaniyu-tonalnosti-russkoyazychnyh-tekstovykh-dannykh> (дата обращения: 15.04.2024).
5. Wankhade, M. A survey on sentiment analysis methods, applications, and challenges [Электронный ресурс] / Wankhade, M., Rao, A.C.S. & Kulkarni, C. – Artificial Intelligence Review, 2022. – 55, 5731–5780. – URL: <https://doi.org/10.1007/s10462-022-10144-1> (дата обращения: 17.04.2024).
6. More than a Feeling: Accuracy and Application of Sentiment Analysis [Электронный ресурс] / Hartmann J., Heitmann M., Siebert C., Schamp C. – International Journal of Research in Marketing, 2023. – Vol. 40, 1. – Pp. 75-87. – ISSN 0167-8116. – URL: <https://doi.org/10.1016/j.ijresmar.2022.05.005> (дата обращения: 17.04.2024).
7. Language models are few-shot learners / Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A. and Agarwal S. – Advances in neural information processing systems. 2020. – No. 33. – Pp. 1877-1901.
8. Краснов, Ф.В. Анализ тональности текста научно-практических статей по нефтегазовой тематике с помощью искусственных нейронных сетей [Электронный ресурс] / Краснов Ф.В. – Вестник Евразийской науки, 2018. – №3. – URL: <https://esj.today/PDF/43ITVN318.pdf> (дата обращения: 15.04.2024).
9. ИТМО: Рекуррентные нейронные сети [Электронный ресурс] / Викиконспекты. – 2024. – URL: [https://neerc.ifmo.ru/wiki/index.php?title=Рекуррентные\\_нейронные\\_сети](https://neerc.ifmo.ru/wiki/index.php?title=Рекуррентные_нейронные_сети) (дата обращения: 16.04.2024).
10. C. Olah. Understanding LSTM Networks. [Электронный ресурс] / C. Olah. – 2015. – URL: <http://colah.github.io/posts/2015-08-Understanding-LSTMs>.
11. A Comparison of Commercial Sentiment Analysis Services [Электронный ресурс] / Ermakova T., Fabian B., Golimblevskaia E., Henke M. – SN COMPUT. SCI., 2023. – Vol. 4, 477. – URL: <https://doi.org/10.1007/s42979-023-01886-y> (дата обращения: 15.04.2024).
12. Brand Analytics : официальный сайт. – Москва, 2012. – URL: <https://brandanalytics.ru/brandanalytics/> (дата обращения: 16.04.2024).
13. Айтеко : официальный сайт. – Москва. — URL: <https://www.i-teco.ru/> (дата обращения: 16.04.2024).
14. Ain Balqis Md Nor Basmmi. Comparison of Web Services for Sentiment Analysis in Social Networking Sites / Ain Balqis Md Nor Basmmi, Shahliza Abd Halim, Nor Azizah Saadon. – IOP Conference Series: Materials Science and Engineering, 2020. – 884. 012063. – DOI 10.1088/1757-899X/884/1/012063.
15. MonkeyLearn : официальный сайт. – 2024. – URL: <https://monkeylearn.com/> (дата обращения: 16.04.2024).
16. The Digital Home of Dr. Daniel Soper : официальный сайт. – 2004. – URL: <https://www.danielsoper.com/default.aspx> (дата обращения: 16.04.2024).
17. Cui, J. Survey on sentiment analysis: evolution of research methods and topics [Электронный ресурс] / Cui, J., Wang, Z., Ho, S. B., & Cambria, E. – Artificial intelligence review, 2023. – Pp. 1–42. – URL: <https://doi.org/10.1007/s10462-022-10386-z> (дата обращения: 17.04.2024).