

УДК 004.048:519.767.6
DOI 10.17513/snt.40386

ОПРЕДЕЛЕНИЕ ТОНАЛЬНОСТИ ТЕКСТОВ МЕТОДАМИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ: АНАЛИЗ ОТЗЫВОВ О ТОРГОВЫХ СЕТЯХ

Ванюлин А.Н., Алексеева Н.Р.

ФГБОУ ВО «Чувашский государственный университет им. И.Н. Ульянова»,
Чебоксары, e-mail: alexis-04@mail.ru

Статья посвящена анализу тональности текста (sentiment analysis) – области компьютерной лингвистики и интеллектуального анализа текста, направленной на извлечение субъективных мнений и эмоций из текстов. Целью исследования является выявление возможностей стандартных методов компьютерной лингвистики для определения тональности отзывов о работе торговых сетей. Основной задачей анализа тональности текста служит определение того, является ли данный текст позитивным, негативным или нейтральным. Рассматриваются различные названия и подходы к данной технологии, такие как мониторинг брендов и анализ социальных медиа, которые получили значительное развитие в последние годы благодаря росту популярности Интернета. В исследовании акцентируется внимание на методах классификации, включая подходы на основе правил и машинного обучения, где последняя категория охватывает современные методы глубокого обучения. Также представлено описание объекта исследования – массива отзывов о торговых сетях – с указанием на методику обработки и анализа данных. Одним из основных результатов является вывод, что необходима тщательная предварительная обработка данных с целью отсеивания дубликатов отзывов. В результате этого с помощью классических методов можно получить правильную классификацию отдельных предложений отзывов с точностью до 72%. Используемый в публикации метод группировки отзывов по соотношению положительных, отрицательных и нейтральных предложений позволяет довести точности определения тональности отзывов до 90%.

Ключевые слова: анализ тональности, обработка естественного языка, компьютерная лингвистика, метод «Bag of Words»

DETERMINATION OF THE TONALITY OF TEXTS BY COMPUTATIONAL LINGUISTICS METHODS: ANALYZING REVIEWS ABOUT RETAIL CHAINS

Vanyulin A.N., Alekseeva N.R.

The Ulianov Chuvash State University, Cheboksary, e-mail: alexis-04@mail.ru

The article is devoted to the analysis of text tonality (sentiment analysis), a field of computational linguistics and text mining aimed at extracting subjective opinions and emotions from texts. The purpose of the study is to identify the possibilities of standard computational linguistics methods for determining the tone of reviews about the work of retail chains. The main task of analyzing the tonality of a text is to determine whether a given text is positive, negative, or neutral. Various names and approaches to this technology are being considered, such as brand monitoring and social media analysis, which have developed significantly in recent years due to the growing popularity of the Internet. The study focuses on classification methods, including rule-based approaches and machine learning, where the latter category covers modern deep learning methods. It also provides a description of the research object – an array of reviews about retail chains, indicating the methods of data processing and analysis. One of the main results is that careful pre-processing of the data is necessary in order to filter out duplicate reviews. As a result, using classical methods, it was possible to obtain the correct classification of individual feedback proposals with an accuracy of 72%. The method of grouping reviews by the ratio of positive, negative and neutral sentences used in the publication makes it possible to increase the accuracy of determining the tone of reviews to 90%.

Keywords: tonality analysis, natural language processing, computational linguistics, «Bag of Words» method

Введение

Автоматический анализ тональности текста представляется одной из активно развивающихся областей компьютерной лингвистики и интеллектуального анализа текста. Результатом анализа является извлечение из него мнений и эмоций по отношению к обрабатываемому тексту.

Попытки подобного анализа начались практически сразу же с появлением систем автоматического анализа текстов. Особое развитие этот вид анализа получил в по-

следние годы в связи с распространением интернет-технологий, в частности с ростом популярности социальных сетей.

Основной задачей в анализе тональности является присвоение данному документу некоторой эмоциональной характеристики (например, является ли описанное в документе мнение позитивным, негативным или нейтральным).

Автоматическое распознавание тональности текстов находит широкое применение в различных сферах деятельности человека [1, с. 127–179].

1. Анализ отзывов о работе торговых сетей. Результаты анализа необходимы для определения предпочтений потребителей по разным видам продукции, степени удовлетворенности качеством услуг и т.д.

2. Сфера финансовых услуг. По характеру высказанных мнений и настроений определяются предпочтения пользователей по выбору банков, кредитных организаций, возможности будущих отношений с определенными финансовыми организациями.

3. Определение тенденций спроса и предложений различных торговых брендов. При этом анализируются отзывы о ценах и качестве товаров с выдачей последующих рекомендаций о возможности или необходимости их покупки.

4. Анализ новостей. Результатом анализа, как правило, является консолидированное мнение относительно произошедших общественных или политических событий.

5. Политика. Анализируются данные об отношении населения к тем или иным политическим деятелям или политическим событиям. Результаты анализа необходимы для организации предвыборных кампаний как отдельных партий, так и конкретных кандидатов.

6. Социология. Анализоваться могут как данные из социальных сетей, так и данные с сайтов, содержащих статистическую информацию (например, демографическую структуру страны (региона, области), уровень потребления различных видов товаров, услуг и продуктов и т.д.).

7. Извлечение несистемной информации. В первую очередь анализируются данные социальных сетей на предмет тенденций общественного мнения относительно общественных и политических процессов и событий.

8. Обратная связь с пользователями. Производится анализ данных социальных сетей, но в первую очередь анализируются запросы и отзывы пользователей о работе государственных и коммерческих структур.

9. Государственная и общественная безопасность. Производится анализ содержания коммерческих и государственных сайтов и частных блогов на предмет антигосударственной или антиобщественной деятельности – распространение наркотиков, экстремистские высказывания, дискредитация государственной политики и т.д.

Все технологии анализа тональности можно разделить на три группы [2].

Первая группа – подходы на основе правил [3–5]. Их характерной особенностью является использование созданных вручную эмоционально размеченных словарей и правил использования данных словарей.

Эти правила определяют класс текста на основе эмоциональных ключевых слов и их совместного использования с другими ключевыми словами.

Вторая группа – подходы на основе машинного обучения [6, 7]. Исторически первыми алгоритмами являются: наивный байесовский классификатор, деревья решений, логистическая регрессия и метод опорных векторов [8, с. 126–130]. В последнее время все чаще применяются методы глубокого обучения, в которых используются сверточные и рекуррентные нейросети, а также методы переноса обучения.

Третья группа – методы анализа на основе на NLP-технологий [9–11].

В зависимости от целей исследования возможны различные варианты классификации текстов. В самом простом случае задача сводится к бинарной классификации, т.е. делению на тексты с положительной и отрицательной тональностью. В зависимости от специфики текста может быть добавлена категория нейтральных текстов. Отнесение текста к той или иной категории обычно осуществляется с использованием дополнительных словарей типа «хорошие» и «плохие» слова или слова, связанные с определенной эмоцией, – «счастливые» или «печальные» слова. Для определения уровня эмоциональности могут быть использованы различные числовые шкалы (например, по количеству позитивных и негативных слов). В результате текст может быть охарактеризован некоторой числовой шкалой, например от –10 для полностью негативных до +10 для полностью позитивных.

Цель исследования – выявление возможностей стандартных методов компьютерной лингвистики для определения тональности отзывов о работе торговых сетей.

Материалы и методы исследования

В настоящем исследовании применен метод «Bag of Words». Основные этапы данного метода выглядят следующим образом.

1. Подготовка данных: необходимо собрать и подготовить данные для обучения модели анализу тональности. Это может быть набор текстовых данных с разметкой, где каждое высказывание классифицировано как положительное, отрицательное или нейтральное.

2. Разбиение предложений на отдельные слова.

3. Удаление стоп-слов. К ним относятся местоимения, предлоги и союзы. Они, как правило, не несут никакой эмоциональной окраски и могут быть удалены. Также удаляются междометия, так как они достаточно часто несут эмоциональную окраску,

но в равной степени как положительную, так и отрицательную.

4. Замена слов на базовые синонимы.

5. Извлечение необходимых признаков (например, формирование списков ключевых слов или их частотностей).

Объектом исследования являлся массив отзывов о работе различных торговых сетей [12]. Общее количество отзывов – 1000.

Результаты исследования и их обсуждение

Основные особенности отзывов:

– отзыв состоит из нескольких предложений;

– анализ каждого отзыва показывает, что он может состоять из предложений различной эмоциональной окраски – положительной, отрицательной и нейтральной.

Поскольку степень «положительности» или «отрицательности» оценить очень трудно и эта оценка чисто индивидуальная, то принято решение о простой фиксации эмоциональности каждого предложения в виде –1, 0 или +1.

В итоге общая тональность каждого отзыва определялась как набор из 2 чисел.

Первое число показывает количество предложений с положительной оценкой, деленное на количество предложений в отзыве, второе – количество предложений с отрицательной оценкой, также деленное на количество предложений в отзыве.

Перед началом непосредственной обработки весь массив отзывов был подвергнут ручной разметке. При этом оценивалась эмоциональность каждого предложения отзыва. Поскольку усредненная длина отзывов составляет 6–7 предложений (совсем точно – 6,51), то фактически вручную была определена тональность примерно 6500 предложений.

Исходя из полученных результатов, все отзывы были разбиты на группы согласно таблице 1. Приведенные в таблице 1 цифры показывают достаточное равномерное распределение отзывов по группам, что свидетельствует о достаточности статистики обрабатываемого объема данных. Следует отметить, что группировка произведена в соответствии с поставленной задачей и в общем носит достаточно приближенный характер.

Последующая обработка состояла из следующих этапов.

1. Из всех предложений были удалены так называемые стоп-слова.

2. Произведена замена слов на базовые словоформы. Например, все вариации слова «стол» («столе», «стола», «столами», «столах», «столом» и т.д.) были заменены словом «стол». При этом использовалась база данных (БД) базовых словоформ объемом более 2 млн слов.

3. Произведена замена слов на базовые синонимы. Например, слова «прекрасный», «очаровательный», «прекрасный», «преlestный», «хорошенький» и иные были заменены на слово «красивый». При этом использовалась БД синонимов объемом примерно в 70 000 слов.

После завершения данного этапа были выявлены дубликаты предложений из различных отзывов. При этом дубликатами являлись только положительные отзывы из групп 1 и 4. Из этого следует очевидное предположение: если какое-то предложение в отзыве является дубликатом, то и весь отзыв является дубликатом.

В результате из всего массива данных было удалено примерно 13% отзывов.

Откорректированные данные по составу и структуре отзывов подробно приведены в таблице 2.

Таблица 1

Процентный состав групп отзывов с различным процентным соотношением предложений разной тональности по результатам ручной разметки (x1 – положительная тональность, x2 – отрицательная тональность, x3 – нейтральная тональность)

Группа	Критерий разделения на группы	Содержание в массиве отзывов, %
1	$x1 > x2$ and $x3 \geq 50\%$	21
2	$x1 < x2$ And $x3 \geq 50\%$	22
3	$x1 = x2$ And $x3 \geq 50\%$	5
4	$x1 > x2$ And $x3 < 50\%$	23
5	$x1 < x2$ And $x3 < 50\%$	26
6	$x1 = x2$ And $x3 < 50\%$	3
Всего		100

Примечание: составлено авторами.

Таблица 2

Процентный состав групп отзывов с различным процентным соотношением предложений разной тональности после удаления дубликатов (x_1 – положительная тональность, x_2 – отрицательная тональность, x_3 – нейтральная тональность)

Группа	Критерий разделения на группы	Содержание в массиве отзывов, %
1	$x_1 > x_2$ and $x_3 \geq 50\%$	15
2	$x_1 < x_2$ And $x_3 \geq 50\%$	22
3	$x_1 = x_2$ And $x_3 \geq 50\%$	5
4	$x_1 > x_2$ And $x_3 < 50\%$	16
5	$x_1 < x_2$ And $x_3 < 50\%$	26
6	$x_1 = x_2$ And $x_3 < 50\%$	3
Всего		87

Примечание: составлено авторами.

Таблица 3

Результаты определения тональности текстов на основе сформированных словарей, %

		Автоматически определенная тональность текста		
		Положительная	Отрицательная	Нейтральная
Тональность текста по результатам ручной разметки	Положительная	72	5	23
	Отрицательная	8	77	15
	Нейтральная	16	5	81

Примечание: составлено авторами.

Таблица 4

Совпадение результатов ручной и автоматической разметки отзывов (x_1 – положительная тональность, x_2 – отрицательная тональность, x_3 – нейтральная тональность)

Группа	Критерий разделения на группы	Совпадение, %
1	$x_1 > x_2$ and $x_3 \geq 50\%$	78
2	$x_1 < x_2$ And $x_3 \geq 50\%$	85
3	$x_1 = x_2$ And $x_3 \geq 50\%$	82
4	$x_1 > x_2$ And $x_3 < 50\%$	91
5	$x_1 < x_2$ And $x_3 < 50\%$	88
6	$x_1 = x_2$ And $x_3 < 50\%$	69

Примечание: составлено авторами.

Полученный результат подтверждает тот факт, что часть отзывов пишут менеджеры соответствующих компаний или специально нанятые люди – ревайтеры (rewriters). Кроме того, этот результат показывает, насколько важным и обязательным является этап предварительной подготовки и фильтрации данных. Поэтому процедура выявления и удаления дубликатов должна быть одной из обязательных процедур предварительного отбора данных для дальнейшей автоматической обработки.

Непосредственно для автоматической обработки были выполнены следующие шаги.

1. Из оставшегося массива данных случайным образом было выбрано 80% отзывов и на основе входящих в них предложений были составлены частотные словари используемых в них слов – отдельный словарь для положительных отзывов и отдельный словарь – для отрицательных.

2. Сформированные словари были проверены на пересечение (на наличие одинаковых слов). По результатам этой операции из обоих словарей были удалены одинаковые слова.

6. С помощью полученных словарей была произведена классификация предложений оставшихся необработанных 20% отзывов.

Результаты классификации приведены в таблице 3.

Как следует из данных таблицы 3, минимально правильная классификация предложений имеет место в 72% случаев. Для описываемой задачи данное значение считается достаточно хорошим результатом [13–15].

Полученные данные были использованы для окончательной группировки отзывов по степени тональности. Основные результаты приведены в таблице 4.

Приведенные в таблице 4 результаты показывают очень высокую эффективность применения стандартных методов компьютерной лингвистики для описываемой задачи. Такая эффективность объясняется, во-первых, тем, что использовались словари, характерные для данного типа текстов, и, во-вторых, относительно приближенным разбиением отзывов на группы. Также эффективность проявляется в том, что даже в случае неточности в определении тональности одного или нескольких предложений соотношение положительных и отрицательных предложений в отзывах остается тем же самым.

Заключение

Проведенное исследование подтвердило эффективность применения стандартных методов компьютерной лингвистики для анализа тональности текста и, в частности, отзывов о торговых сетях. Использование метода «Bag of Words» позволило достичь точности классификации до 90%.

Кроме того, использование данного метода позволяет выявлять дубликаты предложений при 100%-ном совпадении по словарному составу. Этот результат показывает, что этап выявления дубликатов должен быть практически обязательным этапом предварительной обработки данных. В дальнейшем перспективным направлением является комбинирование метода «Bag of Words» с более сложными NLP-технологиями, такими как глубокое машинное обучение, для повышения точности анализа и учета контекстуальных особенностей отзывов. Это позволит более точно оценить эмоциональную окраску текстов и расширить возможности применения анализа тональности в маркетинговых исследованиях и управлении репутацией торговых сетей.

Список литературы

1. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ

данных. М.: НИУ ВШЭ, 2017. 269 с. URL: <https://istina.msu.ru/publications/book/73034754/> (дата обращения: 15.03.2025).

2. Сулейманов А.К., Шарипова М.А., Сметанина О.Н., Сазонова Е.Ю., Миронов К.В. Модели и методы анализа тональности в текстах на башкирском языке. Моделирование, оптимизация и информационные технологии. 2020. Т. 8. № 3. URL: <https://moitvivi.ru/ru/journal/pdf?id=831> (дата обращения: 15.03.2025).

3. Лукашевич Н.В. Автоматический анализ тональности текстов: проблемы и методы // Интеллектуальные системы. Теория и приложения. 2022. Т. 26. № 1. С. 50–61. URL: http://intsysmagazine.ru/issue_26-1 (дата обращения: 12.03.2025).

4. Богданова Т.Ф., Бойчук Е.И. Основные теоретические проблемы, связанные с понятием тональности текста // Верхневолжский филологический вестник. 2020. № 4 (23). С. 136–141. DOI: 10.20323/2499-9679-2020-4-23-136-141.

5. Гаршина В.В., Калабухов К.С., Степанцов В.А., Смотров С.В. Разработка системы анализа тональности текстовой информации // Вестник ВГУ, серия: Системный анализ и информационные технологии. 2017. № 3. С. 185–194. URL: <http://www.vestnik.vsu.ru/pdf/analiz/2017/03/2017-03-21.pdf> (дата обращения: 15.03.2025).

6. Бужин Д.С., Азизов М.Т. Алгоритмы машинного обучения для анализа тональности высказываний // Вестник Российского нового университета. 2022. № 2. С. 129–139. DOI: 10.18137/RNU.V9i87.22.02.P.129.

7. Гималетдинова Г.К., Довтаева Э.Х. Сентимент-анализ читательского комментария: автоматизированная vs ручная обработка текста // Ученые записки Казанского университета. Серия гуманитарные науки. 2021. Т. 163. Кн. 1. DOI: 10.26907/2541-7738.2021.1.

8. Николаев И.С., Митренина О.В., Ландо Т.М. Прикладная и компьютерная лингвистика: коллектив. моногр. Москва: URSS, 2016. 315 с. URL: <https://dokumen.pub/53d0eadb1d981309e7c63df8651a2d90.html> (дата обращения: 17.03.2025).

9. Брунова Е.Г., Бидуля Ю.В. Клиент всегда прав: анализ тональности текста в отзывах о качестве банковского обслуживания // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanities. 2017. Т. 3. № 1. С. 72–89. DOI: 10.21684/2411-197X-2017-3-1-72-89.

10. Гималетдинова Г.К., Довтаева Э.Х. Сентимент-анализ читательского интернет-комментария к политическому тексту // Политическая лингвистика. 2020. №1 (79). DOI: 10.26170/rl20-01-05.

11. Колмогорова А.В., Калинин А.А., Маликова А.В. Типология и комбинаторика вербальных маркеров различных эмоциональных тональностей в интернет-текстах на русском языке // Вестник Томского государственного университета. 2019. № 448. С. 48–58. DOI: 10.17223/15617793/448/6.

12. Отзывы об интернет-магазинах. URL: <https://otzyvshops.com> (дата обращения: 15.03.2025).

13. Рубцова Ю.В. Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015. № 1 (109). С. 72–78. DOI: 10.15827/0236-235X.109.072-078.

14. Лыченко Н.М., Сорокова А.В. Сравнение эффективности методов векторного представления слов для определения тональности текстов // Математические структуры и моделирование. 2019. № 4 (52). С. 97–110. DOI: 10.24147/2222-8772.2019.4.97-110.

15. Baimakhanbetov Mukhit M., Nurumov Kaidar S., Ospanova Ulzhan A., Buldybayev Timur K., Akoyeva Inessa G. The effect of the ADASYN method on Widespread metrics of machine learning Efficiency // Современные информационные технологии и ИТ-образование. 2019. № 2. С. 290–297. DOI: 10.25559/SITITO.15.201902.290-297.