УДК 004.85
doi:10.18720/SPBPU/2/id24-202

Гальцева Татьяна Вячеславовна [1],
студент магистратуры;
Нестеров Сергей Александрович [2],
доцент, канд. техн. наук, доцент

КЛАССИФИКАЦИЯ И ОПРЕДЕЛЕНИЕ ТОНАЛЬНОСТИ ТЕКСТОВ, ПУБЛИКУЕМЫХ В СЕТИ ИНТЕРНЕТ

[1, 2] Россия, Санкт-Петербург,
Санкт-Петербургский политехнический университет Петра Великого;
[1] zaharova.tv@edu.spbstu.ru, [2] nesterov@spbstu.ru

Аннотация. Работа посвящена исследованию различных методов классификации и определения тональности текстов, публикуемых в сети Интернет, а также разработке эффективных подходов для решения данной задачи на основе современных методов машинного обучения и предобработки текстовых данных. В итоге, на основании проведенных исследований, решена задача классификации и определения тональности текстов, которые могут быть использованы для автоматической обработки и анализа больших объемов данных из сети Интернет.

Ключевые слова: классификация, определение тональности, машинное обучение, нейронные сети, предобработка текстов, векторизация.

Tatyana V. Galtseva [1],
BSc, Master Student;
Sergey A. Nesterov [2],
Candidate of Technical Sciences (PhD), Associate Professor

CLASSIFICATION AND SENTIMENT ANALYSIS OF TEXTS PUBLISHED ON THE INTERNET

[1, 2] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia;
[1] zaharova.tv@edu.spbstu.ru, [2] nesterov@spbstu.ru

Abstract. The work is devoted to the study of various methods of classification and determination of the tonality of texts published in the Internet, as well as the development of effective approaches to solve this problem based on modern methods of machine learning and text data preprocessing. As a result, based on the conducted research, the problem of classifying and determining the tonality of texts that can be used for automatic processing and analysis of large amounts of data from the Internet.

Keywords: classification, sentiment analysis, machine learning, neural networks, text preprocessing, vectorization.

Введение

В настоящее время все больше людей используют Интернет для поиска информации о товарах, услугах и мнениях других пользователей.

Классификация и определение тональности текстов позволяют выделить и оценить отзывы, комментарии и другие формы обратной связи пользователей в Интернет по категориям и выявить, являются ли они положительными или отрицательными. Эта информация может быть использована для принятия решений о дальнейшей стратегии развития бизнеса, мониторинге репутации, а также для лучшего понимания потребностей и предпочтений клиентов.

В работе рассматриваются методы классификации и определения тональности текстов, публикуемых в сети Интернет. Для этого будут использованы методы машинного обучения.

1. Постановка задачи

Пусть существует описание документа $d \in X$, где X — векторное пространство документов, и фиксированный набор классов $C = \{c_1, c_2, \ldots, c_m\}$. Из обучающей выборки (множества документов с заранее известными классами) $D = \{\langle d, c\rangle | \langle d, c\rangle \in X \times C\}$ с помощью метода обучения G необходимо получить классифицирующую функцию $G(D) = \gamma$, которая отображает документы в классы [1]:

$$\gamma : X \to C. \quad (1)$$

Целью экспериментальной части работы является классификация отзывов, оставленных пользователями на маркетплейсах, по категориям и определение их тональности.

Для достижения общей цели, связанной с классификацией и определением тональности отзывов, решались следующие задачи:

1) Классификация отзывов.

Цель: определить категорию отзыва.

Входные данные: отзывы пользователей с маркетплейса.

Выходные данные: для каждого отзыва — класс, к которому он относится (цена, качество, удобство использования, качество обслуживания).

2) Определение тональности.

Цель: определить, отзыв является положительным или отрицательным.

Входные данные: отзывы пользователей с маркетплейса.

Выходные данные: для каждого отзыва — его тональность (положительная или отрицательная).

Тогда общей целью будет являться объединение результатов классификации и определения тональности для каждого отзыва.

Для решения данной задачи будут использоваться методы машинного обучения, такие как классификация текстовых данных и анализ тональности. Необходимо провести предобработку данных, включающую в себя удаление стоп-слов, лемматизацию, удаление пунктуации и приведение всех слов к нижнему регистру.

Дополнительная информация о классах:

– Отзывы, которые относятся к классу «цена», будут содержать информацию о стоимости продукта, скидках или акциях.

– Отзывы, которые относятся к классу «качество», будут содержать информацию о качестве продукта, его характеристиках, материалах, из которых он изготовлен, и т. д.

– Отзывы, которые относятся к классу «удобство использования», будут содержать информацию о простоте использования продукта, его функциональности или удобстве хранения.

– Отзывы, которые относятся к классу «качество обслуживания», будут содержать информацию о работе маркетплейса, такую как доставка, оплата или процесс возврата товара.

Результатом работы алгоритма будет определение класса и тональности набора отзывов, что позволит продавцу оценить свой продукт и работу маркетплейса, увидеть, где можно что-то улучшить в своей рабо-

те, а также выделить негативные отзывы, не относящиеся к продукту, а связанные с работой маркетплейса.

2. Выбор технологий для сбора данных

Данные для исследования были получены с помощью парсинга сайта.

Парсинг сайтов — это процесс извлечения информации из веб-страниц. Этот процесс может включать в себя несколько этапов, таких как: скачивание HTML-кода страницы, извлечение необходимой информации из кода, а также сохранение или анализ полученной информации.

Для скачивания HTML-кода страницы можно использовать различные библиотеки и фреймворки, такие как BeautifulSoup, Scrapy, Selenium и другие. Эти библиотеки позволяют отправлять HTTP-запросы к веб-странице и получать ответ в виде HTML-кода.

В работе использовалась библиотека Selenium. Данные собирались с маркет-плейса wildberries.ru, далее размечались вручную для обучения и тестирования. В результате получилось две выборки:

1. Данные, для обучения модели определять категорию отзыва: 1039 строк. Отзывов, относящихся к категории «Удобство использования» — 251, «Качество» — 278, «Цена» — 257, «Качество обслуживания» — 252. Отзывы, которые не относились ни к одной из трех других категорий, были определены в категорию «Качество».

2. Данные, для обучения модели определять тональность отзыва: 2000 строк. Количество отзывов, имеющих отрицательную тональность — 1000, положительную тональность — 1000.

3. Предобработка текстовых данных

Для корректной работы алгоритмов классификации необходимо провести нормализацию и векторизацию текстовых данных. Для нормализации были использованы различные модули и библиотеки языка Python:

– pandas — библиотека для работы с данными в таблицах, которая используется для чтения и записи данных из файлов формата CSV, а также для обработки данных в таблице;

– tqdm — библиотека для отображения прогресса выполнения итераций цикла;

– nltk — библиотека для работы с естественным языком, которая используется для удаления стоп-слов и токенизации текста;

– string — модуль для работы со строками, который используется для удаления знаков пунктуации из текста;

– re — модуль для работы с регулярными выражениями, который используется для удаления множественных пробелов в тексте;

– pymystem3 — библиотека для лемматизации русских слов, которая используется для лемматизации текста;

– SnowballStemmer — класс для стемминга русских слов, который используется для стемминга текста;

– stopwords — корпус стоп-слов для русского языка из библиотеки nltk, который используется для удаления стоп-слов из текста.

Для векторизации текста используется Bag of Words (сокр. BoW, в переводе с англ. «мешок слов») — это один из наиболее простых и широко используемых методов векторизации текста. Модель мешка слов представляет упрощенное представление текстовых документов, основанное на учете частоты слов в документе без учета их порядка или контекста.

В модели BoW каждый текстовый документ рассматривается как набор слов, при этом игнорируется их порядок. При создании модели формируется словарь, который отображает каждое слово в уникальный идентификатор. Затем для каждого документа создается вектор, в котором каждый элемент соответствует количеству вхождений соответствующего слова из словаря [2].

4. Методы классификации текстов

В представляемой работе рассмотрены только методы обучения с учителем, поскольку классы текстов известны заранее. Ниже перечислены алгоритмы классификации, использовавшиеся в работе:

– Упрощенный алгоритм Байеса (англ. Naive Bayes) — алгоритм классификации, основанный на вычислении условной вероятности значений прогнозируемых атрибутов. При этом предполагается, что входные атрибуты являются независимыми и определен хотя бы один выходной атрибут;

– Метод опорных векторов (англ. Support Vector Machine, SVM) — цель метода заключается в нахождении оптимальной разделяющей гиперплоскости среди всех возможных гиперплоскостей пространства, отделяющих два класса обучающих примеров друг от друга, такой гиперплоскости, расстояния от которой до ближайших векторов обоих классов равны [3];

– Логистическая регрессия (англ. Logistic Regression) — основная идея логистической регрессии заключается в построении модели, которая предсказывает вероятность отнесения объекта к определенному классу [3];

– Деревья решений (англ. Decision Tree) — семейство алгоритмов, позволяющих сформировать правила классификации в виде иерархической (древовидной структуры) [4].

5. Полученные результаты

В результате получены наборы данных, которые были размечены вручную по категориям отзывов, а также по тональностям. Эти данные использовались при обучении и тестировании моделей.

В таблице 1 представлены метрики бинарной классификации (в работе — классификация по тональностям), которые показывают результаты работы методов классификации. Каждый метод оценивается по таким метрикам, как точность, полнота, F1-мера, а также AUC-ROC.

Таблица 1

Метрики бинарной классификации

Метод классификации	Точность	Полнота	F1-мера	AUC-ROC
Naive Bayes Classifier	**0.93**	**0.93**	**0.93**	**0.98**
Logistic Regression	0.92	0.92	0.92	0.96
Support Vector Machine	0.91	0.91	0.91	0.94
Decision Tree	0.84	0.84	0.84	0.84

Исходя из анализа таблицы, можно сделать вывод, что Naive Bayes Classifier является наиболее эффективными методом классификации на основе предоставленных в работе данных для обучения модели определять тональность отзыва.

В таблице 2 представлены метрики многоклассовой классификации (в работе — классификация по категориям), которые показывают результаты работы методов классификации. Каждый метод оценивается по таким метрикам, как точность, полнота и F1-мера.

Таблица 2

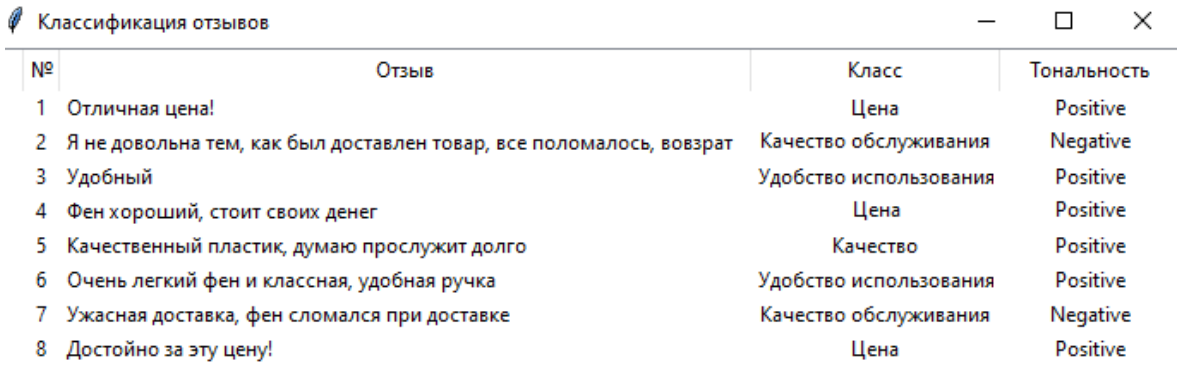
Метрики многоклассовой классификации

Метод классификации	Точность	Полнота	F1-мера
Naive Bayes Classifier	0.82	0.82	0.82
Logistic Regression	**0.87**	**0.87**	**0.87**
Support Vector Machine	0.83	0.83	0.83
Decision Tree	0.80	0.80	0.80

Исходя из анализа таблицы, можно сделать вывод, что метод Logistic Regression является наиболее эффективными методом классификации на основе предоставленных данных для обучения модели определять категорию отзыва на товар.

6. Тестирование модели на реальных данных

С помощью парсинга были собраны отзывы с Интернет-магазина www.wildberries.ru. Результат работы программы показан на рисунке 1.



№	Отзыв	Класс	Тональность
1	Отличная цена!	Цена	Positive
2	Я не довольна тем, как был доставлен товар, все поломалось, вовзрат	Качество обслуживания	Negative
3	Удобный	Удобство использования	Positive
4	Фен хороший, стоит своих денег	Цена	Positive
5	Качественный пластик, думаю прослужит долго	Качество	Positive
6	Очень легкий фен и классная, удобная ручка	Удобство использования	Positive
7	Ужасная доставка, фен сломался при доставке	Качество обслуживания	Negative
8	Достойно за эту цену!	Цена	Positive

Рис. 1. Результат работы программы на реальных отзывах

Для визуализации результатов были построены круговые диаграммы, показывающие процент положительных и отрицательных отзывов для каждого класса (рис. 2).

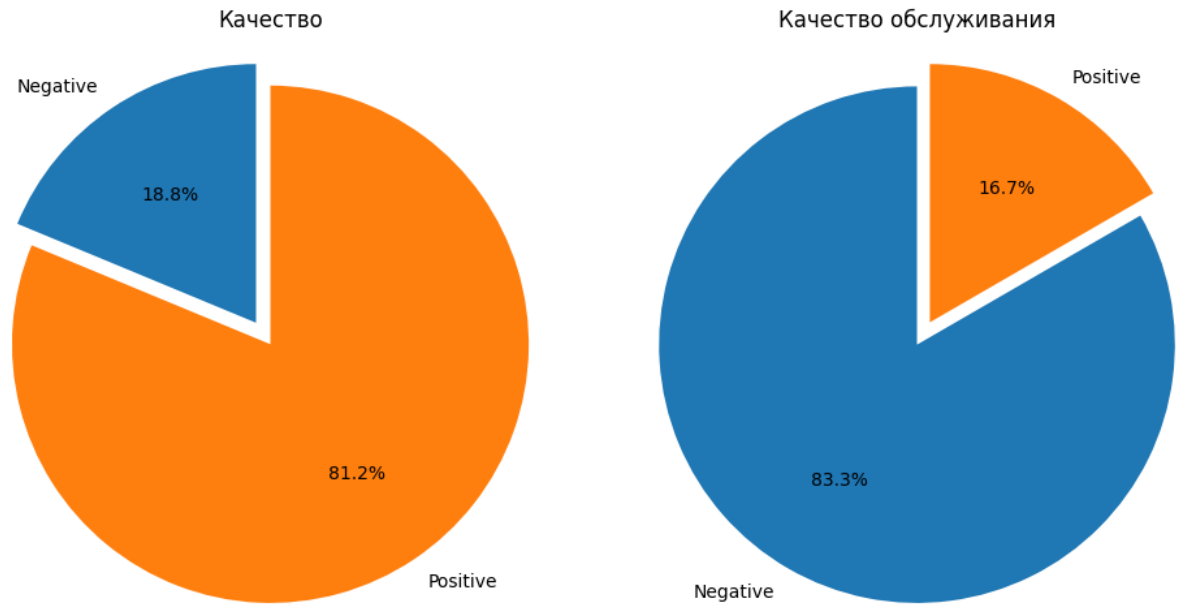


Рис. 2. Круговые диаграммы для каждого класса

Заключение

В ходе работы были получены следующие результаты:

1) Разработан модуль сбора данных. Подготовлены наборы данных для обучения и тестирования.
2) Изучены и применены различные методы предварительной обработки текста, включая токенизацию, удаление стоп-слов, лемматизацию и стемминг. Эти методы позволяют стандартизировать текстовые данные и повысить качество классификации.

3) Были обучены модели, способные определять тональность отзывов и классифицировать их по категориям.
4) Проведен сравнительный анализ точности различных методов классификации и определения тональностей. Наилучшие результаты показали модели на основе упрощенного алгоритма Байеса и логистической регрессии.

Полученные результаты в классификации и определении тональности текстов могут быть использованы в различных областях. Вот некоторые из возможных вариантов использования:

1) Анализ отзывов и мнений пользователей: полученные методы можно использовать для классификации и анализа отзывов в различных областях, таких как интернет-магазины, бронирование гостиниц, рестораны и т. д. Это позволяет компаниям понять мнения и предпочтения своих клиентов и принять соответствующие меры.
2) Социальный мониторинг: методы классификации текста можно использовать для анализа и отслеживания мнений и настроений в социальных сетях и на других платформах. Это позволяет организациям и исследователям получить представление о реакции на новые продукты, события, политические решения и т. д.
3) Обработка клиентских запросов: полученные методы могут применяться для автоматической классификации и анализа клиентских запросов, писем, сообщений в чатах и т. д. Это может помочь компаниям расставлять приоритеты и повышать качество обслуживания.

Список литературы

1. Романов А. С. и др. Анализ тональности текста с использованием методов машинного обучения. [Электронный ресурс]. – URL: http://ceur-ws.org/Vol-2233/Paper_8.pdf (дата обращения: 22.10.2023).

2. Kowsari K., Jafari Meimandi K., Heidarysafa M. et. al. Text classification algorithms: A survey // Information. – 2019. – Vol. 10. No 4. – P. 150

3. Баев Н.О. Использование метода опорных векторов в задачах классификации // Международный журнал информационных технологий и энергоэффективности. – 2017. – Т.2. – № 2(4). – С. 17–21.

4. Нестеров С.А. Базы данных. Интеллектуальный анализ данных: учеб. пособие. – СПб.: Изд-во Политехн. ун-та, 2011.