АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТОВОЙ ИНФОРМАЦИИ НА ОСНОВЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

А. Д. Журлов

Саратовский национальный исследовательский государственный университет им. Н.Г. Чернышевского, Россия
E-mail: olomaster01@gmail.com

В данной работе исследуются возможности применения методов машинного обучения для анализа тональности текстовой информации. Описывается процесс подготовки данных, который включает сбор данных, очистку от шума, токенизацию, нормализацию и удаление стоп-слов. На полученном наборе данных проводятся эксперименты с использованием 4 методов классического машинного обучения: метод опорных векторов (SVM), метод k-ближайших соседей (KNN), логистическая регрессия и метод градиентного бустинга.

Результаты показали, что наилучшие показатели достигнуты с моделью логистической регрессии, точность которой составила 70,81% и метрика F1 - мера (weighted) - 50%.

SENTIMENT ANALYSIS OF TEXT INFORMATION BASED ON MACHINE LEARNING ALGORITHMS

A. D. Zhurlov

This work explores the possibilities of using machine learning methods to analyze the sentiment of text information. The data preparation process is described, which includes data collection, noise removal, tokenization, normalization, and stop word removal. Experiments are conducted on the resulting dataset using 4 classical machine learning methods: support vector machine (SVM), k-nearest neighbors (KNN), logistic regression, and gradient boosting.

The results showed that the best scores were achieved with the logistic regression model, the accuracy of which was 70,81% and the F1 - measure metric (weighted) was 50%.

При выполнении работы использовались теоретические материалы ([1]-[4]).

В современном мире тема анализа тональности текста и речи крайне актуальна. Технология распознавания эмоций активно внедряется в системы искусственного интеллекта, поскольку она позволяет более точно интерпретировать человеческие чувства и адаптировать взаимодействие ИИ в зависимости от настроения пользователя. Это добавляет больше вариативности ответам и улучшает их качество. Также инструменты по определению тональности текстов чрезвычайно полезны для владельцев развлекательных сервисов и других подобных бизнесов, так как отзывы клиентов могут указать направление, в котором компании будет выгоднее развиваться. Понимание эмоциональной окраски отзывов может помочь в выявлении сильных и слабых сторон продукта, что способствует повышению качества услуг и удовлетворенности клиентов.

В данном исследовании проводилась оценка тональности комментариев пользователей сайта «Кино Mail».

В качестве тестовых данных был выбран датасет, содержащий в себе отзывы клиентов множества отелей и гостиниц, т.к. в этой сфере большинство сообщений пользователей содержат в себе множество наборов эмоций. Тестовая выборка данных составляла 30% из общего числа отзывов.

Для оценки тональности размеченных данных использовались классические методы машинного обучения из пакета scikit-learn [5].

При предобработке отзывов использовался лемматизатор Mystem.

Для таких методов машинного обучения, как логистическая регрессия (Logistic Regression), метод опорных векторов (SVM), градиентный бустинг (Gradient Boosting) и метод k-ближайших соседей (KNN) были использованы представления слов в виде мешка слов (Bag of Words).

Для оценки качества классификации использовалась метрика F1-мера с взвешенным усреднением (weighted). Выбор именно этой метрики для оценки тональности текста объясняется тем, что имеющийся набор данных не сбалансирован. Лучший результат по F1-мере был получен для набора BagOfWords + Logistic Regression (мешок слов + метод логистической регрессии) на лемматизированном тексте, при этом весовая F1-мера составила 50%. Полученные результаты представлены в таблице ниже (лучший результат выделен).

Результаты экспериментов по применению методов машинного обучения для оценки тональности текста

Модель	Accuracy, %	F1 weighted, %
BoW + Logistic Regression	**70,81**	**50,00**
BoW + GradientBoostingClassifier	70,38	20,00
BoW + SVM	74,31	33,33
BoW + KNN	62,40	33,33

В будущем планируется провести эксперименты на сбалансированных данных, а также продолжить исследование с использованием большего количества моделей машинного обучения для получения лучшего результата.

СПИСОК ЛИТЕРАТУРЫ

1. *Фридман Дж. Х., Тибширани Р. Х., Хастие Т.* Чистый машинный анализ: Принципы и приложения. – М. : Гаудеамус, 2014. 450 с.

2. *Бишоп К. М.* Машинное обучение: Исходный текст. М. : ДМК Пресс, 2013. 512 с.

3. *Bird S., Klein E., Loper E*. Natural Language Processing with Python. O'Reilly Media, 2009. 504 p.

4. *Бишоп К. М.* Pattern Recognition and Machine Learning. - Springer, 2006. 738 p.

5. Scikit-learn: Machine Learning in Python [Электронный ресурс]. URL: https://scikit-learn.org/stable/ (дата обращения: 10.08.2024).