УДК 004.855.5

АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТОВ ИЗ СОЦИАЛЬНЫХ МЕДИА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Гончаров А. В.
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к.т.н., Махныткина О.В.
(Университет ИТМО, г. Санкт-Петербург)

Введение. Массовое увеличение пользовательского контента, в том числе содержащего мнения и отзывы о различных товарах, услугах, проблемах, событиях способствует росту востребованности автоматической оценки тональности текстов. Оценка тональности текста — это междисциплинарная область, включающая обработку естественного языка и машинное обучение. В работе предлагается исследование методов предварительной обработки текстовых данных, извлечения признаков и классификации для оценки тональности текстов на русском языке.

Основная часть. Цель работы - исследование методов машинного обучения для оценки тональности текста. Оценка качества моделей проводилась на основе таких метрик как точность, полнота, F1-мера. Для оценки тональности текстов использовался новостной датасет с платформы Kaggle, содержащий 2000 сообщений в тестовом множестве и 8000 сообщений в обучающем множестве. Разметка сообщений проводилась по трем классам: позитивные, негативные, нейтральные.
Для оценки тональности текстов из социальных медиа необходимо учитывать специфику языка и использование сокращений, которые могут повлиять на точность анализа. Для повышение точности классификации, в качестве инструментов предварительной обработки было использовано: приведение слов к нижнему регистру, обработка знаков пунктуации, токенизация и удаление стоп слов. В работе также была применена нормализация слов и извлечение признаков с использованием модели fastext [1]. Для классификации сообщений были использованы как классические методы машинного обучения (k-средних, деревья решений, метод опорных векторов), так и нейронные сети [2].

Выводы. Результатом работы является сравнительный анализ методов предварительной обработки тестовых данных и машинного обучения для оценки тональности текстов на русском языке.

Список использованных источников

1. Двойникова А. А. , Карпов А. А. Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных – Обработка информации и управление, doi:10.31799/1684-8853-2020-4-20-30 , 2020 – стр. 20-30
2. Smetanin S. The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives - IEEE Access, DOI: 10.1109/ACCESS.2020.3002215 2020 – pp. 110693 - 110719