Селиверстов Александр Витальевич,
студент магистратуры 1 курса гр. Ф2-ИСТ-51м,
ФГОБУ ВО «Поволжский государственный университет
телекоммуникаций и информатики»

Захарова Оксана Игоревна, к.т.н, доцент,
ФГОБУ ВО «Поволжский государственный университет
телекоммуникаций и информатики»

ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА: МЕТОДЫ, БИБЛИОТЕКИ PYTHON И МЕХАНИЗМЫ АНАЛИЗА ТОНАЛЬНОСТИ

Аннотация. В данной статье рассматриваются современные подходы к обработке естественного языка (NLP) и инструменты на базе Python, используемые для автоматического анализа текста. В ней описываются основные этапы NLP, включая предварительную обработку, лингвистическую нормализацию и извлечение признаков, а также подробно описываются методы анализа тональности, от методов на основе лексикона до нейронных архитектур на основе трансформеров. Особое внимание уделяется преимуществам контекстных языковых моделей для обработки пользовательского контента на русском языке, такого как отзывы о покупках.

Ключевые слова: NLP, обработка текста, Python, анализ тональности, машинное обучение.

Современная цифровая среда характеризуется постоянным увеличением объема текстовых данных. Отзывы пользователей, комментарии, техническая документация формируют массивы информации, требующие автоматизированный обработки. Одним из ключевых направлений, обеспечивающих такую обработку, является обработка естественного языка (Natural Language Processing - NLP). Этот раздел искусственного интеллекта направлен на извлечение полезной информации из текстов и их структурированный анализ.

Процесс обработки естественного языка включает комплекс методов, обеспечивающих преобразование текста в форму, пригодную для вычислительной обработки. Применение библиотек Python позволяет эффективно реализовать такие процессы в научных, исследовательских и прикладных задачах.

Одним из первичных этапов обработки является сбор и отчистка данных. На этом этапе выполняется удаление шумов, таких как HTML-разметка, эмодзи, дубликаты, а также нормализация регистра. Важным является удаление стоп-слов, однако для русскоязычных текстов требуется учитывать отрицательные конструкции, поскольку выражение «не работает» и «работает» имеет противоположный смысл.

После отчистки выполняется токенизация, стемминг и лемматизация. Лемматизация предпочтительнее для русского языка, поскольку она корректно учитывает сложную морфологию и позволяет получать словарную форму слова. Затем текст преобразуется в числовые представления – TF-IDF, Word2Vec, FastText или контекстные векторы BERT [1].

Ключевым этапом анализа является применение алгоритмов машинного обучения. Традиционные методы включают логистическую регрессию, SVM и наивный байесовский классификатор. Однако наибольшую эффективность демонстрируют модели на основе трансформеров, такие как BERT и его русскоязычные модификации (RuBERT, blinoff/bert-base-russian-uncased) [2]. Благодаря механизмам самовнимания такие модели учитывают контекст каждого слова в предложении, что особенно важно при анализе отзывов.

Применение NLP-технологий в анализе пользовательских отзывов позволяет определять тональность сообщений, выявлять повторяющиеся проблемы и формировать агрегированные характеристики товара. Тексты отзывов часто содержат эмоционально окрашенные конструкции, короткие фразы, разговорные выражения, что требует использования контекстных моделей.

К примеру, такие выражения как товар не плохой и товар плохой имеют противоположные оценки, и только модели контекстного типа способны корректно их различать. Анализ частотных слов в сочетании с определением тональности позволяет формировать целостную картину восприятия продукта покупателям.

Таким образом, обработка естественного языка является ключевым инструментом анализа текстовой информации. Однако практическое применение NLP в реальных проектах предполагает не только использование готовых библиотек, но и комплексную архитектуру системы, включающую сбор данных, их нормализацию, анализ и интеграцию с пользовательским интерфейсом.

Список литературы:

1. Анализ тональности в русскоязычных текстах, часть 1: введение [Электронный ресурс] Режим доступа: https://habr.com/ru/companies/vk/articles/516214/, свободный. – Загл. с экрана.

2. Адаптация глубоких двунаправленных многоязычных трансформеров для русского языка [Электронный ресурс] Режим доступа: https://arxiv.org/abs/1905.07213, свободный. – Загл. с экрана.