УДК 004.912

ОПРЕДЕЛЕНИЕ ТОНАЛЬНОСТИ ТЕКСТА С ПОМОЩЬЮ МЕТОДОВ NLP

***Лумпова Полина Сергеевна**, студентка 3 курса магистратуры института экономики и управления АПК, ФГБОУ ВО РГАУ – МСХА имени К. А. Тимирязева, polinalumpova@gmail.com*

***Научный руководитель – Ульянкин Александр Евгеньевич**, ассистент кафедры статистики и кибернетики института экономики и управления АПК, ФГБОУ ВО РГАУ – МСХА имени К. А. Тимирязева, aeulianckin@rgau-msha.ru*

***Аннотация**. В статье рассматриваются методы анализа клиентских отзывов с использованием технологий обработки естественного языка (NLP). Основное внимание уделяется разработке модели для оценки тональности текстов и выделения ключевых слов. Описываются этапы работы предложенной модели, приводятся результаты точности ее работы. Отмечены вопросы, нуждающиеся в доработке.*

***Ключевые слова**: машинное обучение, обработка естественного языка, тональность, обработка текста, глубинный анализ данных.*

DETERMINING THE TONALITY OF A TEXT USING NLP METHODS

***Lumpova Polina Sergeevna**, 3rd year graduate student of the Institute of Economics and Management of the Agroindustrial Complex, Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, polinalumpova@gmail.com*

***Scientific supervisor – Ulyankin Alexander Evgenyevich**, assistant at the Department of Statistics and Cybernetics of the Institute of Economics and Management of the Agroindustrial Complex, Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, aeulianckin@rgau-msha.ru*

***Annotation.** The article discusses methods for analyzing customer reviews using Natural Language Processing (NLP) technologies. The main focus is on developing a model for sentiment analysis and keyword extraction. The stages of the proposed model's operation are described, and its accuracy results are presented. Issues that require further improvement are highlighted.*

***Key words**: machine learning, natural language processing, sentiment, text processing, data mining.*

Процесс предоставления товаров и услуг потребителям требует систематического сбора обратной связи, представленной в виде оценок и

текстовых отзывов. Однако сам по себе сбор оценочных данных не всегда обеспечивает достаточную информативность для точной интерпретации клиентских отзывов: в случаях, когда услуга состоит из множества компонент, низкая оценка может неявно указывать на неудовлетворённость одним или несколькими аспектами. В подобных ситуациях возникает необходимость в анализе отдельных составляющих [3]. При анализе исключительно текста отзывов возникает проблема обработки значительных объемов неструктурированной информации, что усложняет процесс выделения ключевых элементов и точного определения эмоциональной окраски клиентских оценок. Данная задача требует применения методов обработки естественного языка (Natural Language Processing, NLP), которые способны структурировать текстовые данные и извлекать информативные маркеры тональности [2].

Разрабатываемая модель для оценки тональности отзыва и выделении ключевых слов работает по следующему принципу (Рисунок 1): каждому слову присваивается вес негативной окраски или позитивной, веса всех слов одного отзыва складываются тоже по негативной или позитивной окраске, сравниваются суммы и выдается результат.

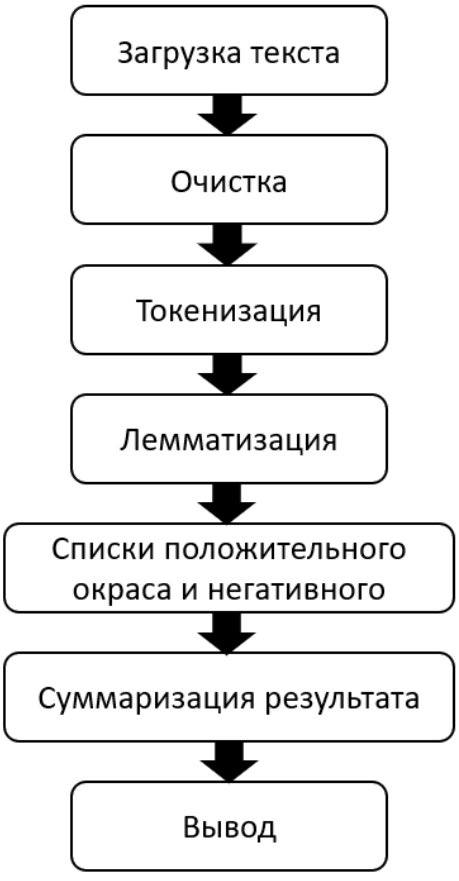

Рисунок 3 – **Процесс работы предлагаемой модели**

Первоначальная очистка текстовых данных включает в себя удаление знаков препинания, так как они не несут смысловой нагрузки, однако требуют затрат времени и ресурсов для обработки. Токенизация представляет собой процесс, в ходе которого исходный текст разбивается на отдельные

предложения, которые, в свою очередь, разделяются на слова [5]. При этом сохраняется порядок слов, что обеспечивает сохранение синтаксической структуры предложений. Лемматизация — это еще один ключевой этап, позволяющий приводить слова к их начальной форме, устраняя склонения и временные формы глаголов [5]. Этот процесс крайне важен, поскольку одни и те же слова могут встречаться в различных формах, что упрощает работу со словарями, используемыми для обучения моделей. Для выполнения этих задач используется библиотека Natasha, обладающая поддержкой русского языка, что выделяет ее на фоне других крупных библиотек для обработки естественного языка [4].

После предварительной обработки данных начинается этап анализа тональности. На этом этапе слова из текста сопоставляются со словами из заранее составленного списка, каждому из которых присваивается вес, отражающий его положительную или отрицательную окраску. Формируется отдельный список для каждого текста, представленный в виде оценок. Суммарные результаты по каждому тексту агрегируются, что позволяет классифицировать общий эмоциональный фон отзыва с выделением позитивных и негативных сегментов [1]. Сохранение синтаксической структуры текста позволяет сегментировать его на отдельные смысловые блоки, что существенно повышает точность анализа и способствует выделению конкретных проблемных областей [4].

Пример того, как может выглядеть визуальный результат анализа текста представлен ниже (Рисунок 2). Зеленым выделены слова с положительной окраской, красным-негативной. Градиент показывает, какое из слов имеет наибольшее значение в том или ином настроении. При этом видно, как модель объединяет в блоки связанные или зависимые слова. Благодаря этому в дальнейшем можно сделать разбивку на темы.

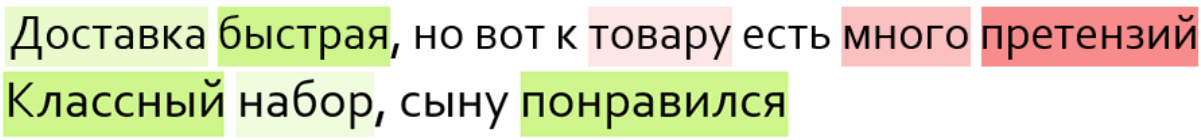

Рисунок 2 – **Пример выделения позитива и негатива в отзывах**

В таблице (Рисунок 3) представлены результаты модели на всем тексте отзыва, без деления его на блоки. Видно, что очень большая ошибка первого рода, когда мы негатив принимаем за позитив. Больше всего это связано с отзывами, в которых люди используют слова с положительной окраской, но в негативном смысле, например, сарказм. При этом, есть улучшение точности модели при делении отзыва на блоки (Рисунок 4). Так отзывы, в которых есть негатив и позитив не смешиваются, а каждый блок анализируется отдельно, имея свое настроение.

	Neg	Pos
True		
Neg	0.72	0.28
Pos	0.16	0.84

Рисунок 3 – **Матрица ошибок модели на всем тексте отзыва**

	Neg	Pos
True		
Neg	0.76	0.24
Pos	0.13	0.87

Рисунок 4 – **Матрица ошибок модели с делением на блоки отзыва**

Таким образом, анализ текстов отзывов с использованием методов обработки естественного языка является эффективным инструментом для глубинного изучения мнений потребителей. Применение NLP-технологий способствует систематизации и структурированию больших объемов текстовой информации, выявлению ключевых тенденций и закономерностей в отзывах клиентов, что, в свою очередь, позволяет улучшить качество обслуживания и удовлетворённость пользователей. Однако наиболее эффективные результаты достигаются при разработке собственных моделей, обученных на уникальных данных. Это позволяет учитывать специфические нюансы сферы, что значительно повышает точность анализа.

Текущая модель нуждается в улучшении, в том числе через снижение ошибки первого рода. Но для этого инструментов для работы с языком и для обучения недостаточно. Необходимо учитывать и общий портрет человека, который его оставил, а также триггеры, которые предшествовали обратной связи от клиента [3]. Для общего анализа настроения клиентов стандартных инструментов NLP достаточно, но для полноценной работы со всеми данными уже необходимо использовать более мощные инструменты глубокого обучения [2].

Библиографический список

1. Alharbi J.R., Alhalabi W.S. Hybrid Approach for Sentiment Analysis of Twitter Posts Using a Dictionary-based Approach and Fuzzy Logic Methods: Study Case on Cloud Service Providers, International Journal on Semantic Web and Information Systems (IJSWIS), 2020, Vol. 16, No. 1, pp. 116-145.

2. Natural Language Processing with Python // NLTK URL: https://www.nltk.org/book/ (дата обращения: 26.10.2024).

3. Майорова Е.В. О сентимент-анализе и перспективах его применения / Майорова Е.В. [Электронный ресурс] // КиберЛенинка: [сайт]. — URL: https://cyberleninka.ru/article/n/o-sentiment-analize-i-perspektivah-ego-primeneniya?ysclid=lolbjlkfom908044470 (дата обращения: 26.10.2024).

4. Проект Natasha. Набор качественных открытых инструментов для обработки естественного русского языка (NLP) // Хабр URL: https://habr.com/ru/articles/516098/ (дата обращения: 26.10.2024).

5. Хобсон Лейн, Ханнес Хапке, Коул Ховард Обработка естественного языка в действии. — СПб.: Питер, 2020. — 576 с.

6. Романова, М. А. Алгоритмы обработки текста / М. А. Романова, Д. Э. Храмов // Материалы международной научно-практической конференции "Тренды развития сельского хозяйства и агрообразования в парадигме Зеленой экономики" : сборник статей, Москва, 14–15 июня 2023 года. – Москва: Российский государственный аграрный университет- Московская сельскохозяйственная академия им. К.А. Тимирязева, 2023. – С. 252-257. – EDN HZDKZR.