# АНАЛИТИЧЕСКИЙ ОБЗОР

## Введение

Онлайн-обзоры имеют большое значение для поддержки принятия потребителями решений о покупке. В настоящее время существует огромное количество платформ, которые позволяют пользователям размещать отзывы о различных товарах, услуги или рассказать о своем опыте работы в той или иной компании; эти обзоры отражают потребительский опыт и снижают неопределенность при онлайн-покупках [1]. Однако часто на популярные товары оставляется огромное количество отзывов и появляется задача ранжирования отзывов по полезности [2].

По результатам исследования Local Consumer Review Survey 2022 все больше и больше потребителей читают отзывы. В 2021 году 77% "всегда" или "регулярно" читали их перед покупкой товара (по сравнению с 60% в 2020 году), в 2022 году эта цифра увеличилась до 98%. 57% утверждают, что они были бы не склонны приобретать товары у продавцов, которые не реагирует на отзывы.

Таким образом анализ отзывов может быть полезен не только потенциальным потребителям, но и владельцам компании для контроля качества своей продукции. Особенно это может быть актуально для компаний, специализирующихся на здравоохранении, техническом обслуживании, выпуску уходовой продукции и косметических средств.

Для ручного анализа отзывов и настроений клиента к различным аспектам товара, услуги или компании в целом может потребоваться большое количество времени, в некоторых случаях даже месяцы. При анализе отзывов могут возникать различные проблемы - отзывы могут быть неоднозначными или бесполезными, к такой информации также необходим особенный подход.

В области анализа тональности текста сейчас актуально использование трансформеров для подробного анализа отзывов и др. текстов, т. е. не просто предоставление рейтинга всего текста, а полный отчет об отношении к различным объектам, например отношение пользователя к самому фильму, к режиссеру или актерам. Это должно позволить строить более точные рекомендательные системы и облегчить контроль качества товаров.

Таким образом **целью работы** стало разработать сервис для подробного анализа отзывов пользователей с использованием трансформеров и их ранжирование по степени полезности.

**Задачами работы** являются:

1. Составить список источников для аналитического обзора
2. Провести анализ существующих исследований и архитектур нейронных сетей для анализа тональности текста
3. Подготовить наборы данных для тренировки, валидации и тестирования нейронной сети, которые можно использовать для дальнейшей работы
4. Разработать архитектуру нейронной сети и протестировать ее
5. Разработать сервис для подробного анализа отзывов пользователей

**Объектом исследования** является – трансформеры.

**Предметом исследования** является – применение трансформеров для эффективного анализа эмоционально окраски отзывов.

## Степень научной изученности и разработанности выбранной темы

Анализ настроений (SA) — это класс вычислительных методов, который автоматически извлекает и обобщает мнения о таком огромном объеме данных, что исследователь в одиночку не может обрабатывать. Автоматизированная обработка текстов также достигла значительного прогресса. Десять лет назад стандартом в анализе текста был подход bag-of-words, который пять лет спустя сменился подходом с использованием нейронных сетей на основе рекуррентных слоев (LSTM, GRU), сейчас рассматриваются стандартные модели с архитектурой transformer. Эти изменения привели к технологическому усложнению подходов и увеличению способности машин понимать язык, в том числе основанный на символах, отличных от латиницы. [3]

Трансформеры [4] — относительно новый тип нейросетей, направленный на решение последовательностей с легкой обработкой дальнодействующих зависимостей. На сегодня это самая продвинутая техника в области обработки естественной речи (NLP). С их помощью можно переводить текст, писать стихи и статьи и даже генерировать компьютерный код. В отличие от рекуррентных нейронных сетей (RNN), трансформеры не обрабатывают последовательности по порядку, поэтому можно ускорить обучение, используя больше графических процессоров. Например, если исходные данные — текст, то им не нужно обрабатывать конец предложения после обработки начала. Благодаря этому такую нейросеть можно распараллелить и обучить значительно быстрее. Также RNN сложно тренировать, они подвержены так называемой проблеме исчезающего/взрывающегося градиента.

Основными компонентами трансформеров являются энкодер и декодер. Энкодер преобразовывает входящую информацию (например, текст) и конвертирует ее в вектор (набор чисел). Декодер, в свою очередь, расшифровывает ее в виде новой последовательности (например, ответ на вопрос) слов на другом языке — смотря для каких целей создавалась нейросеть.

Одним из самых простых примеров использования трансформеров в NLP является языковая модель BERT компании Google, разработанная в 2018 году. Другой пример популярной языковой модели на базе трансформеров — это GPT (Generative Pre-trained Transformer) компании OpenAI.

В таблице 1 представлены наборы данных для анализа тональности русскоязычного текста, которые могут быть использованы в процессе исследования проблемы.

## Анализ первоисточников

В статье The Effect of Online Reviews on Product Sales: A Joint Sentiment-Topic Analysis [5] авторы Xiaolin Li, Chaojiang Wu и Feng Mai утверждают, что созданная ими унифицированная модель joint sentiment-topic model для анализа настроений отзывов достигает сразу двух целей одновременно: она не только суммирует настроения в тексте обзора, но и определяет аспекты продукта, которыми рецензент доволен или к которым относится критически. JST обеспечивает гораздо более богатое представление качественных обзорных данных. Их результаты позволяют исследовать, среди прочего, насколько положительна или отрицательна валентность конкретных характеристик продукта и как они приводят к изменениям в будущих продажах. В своем исследовании они изучали влияние текстовых обзоров и числовых оценок на фактические продажи 312 продуктов для планшетных ПК.

Авторы Peleshchak R. и др. в своей статье Text Tonality Classification Using a Hybrid Convolutional Neural Network with Parallel and Sequential Connections Between Layers [6] предлагают иную архитектуру нейронной сети для обработки настроений – гибридную сверточную нейронную сеть, которая извлекает текстовые объекты, используя параллельно подключенный сверточный блок. Затем нейронная сеть классифицирует объекты и объединяет эти объекты с исходными текстовыми объектами. Модель предлагаемой нейронной сети способна изучать как локальные, так и глобальные особенности коротких текстов и имеет меньшее время конвергенции и вычислительный ресурс по сравнению с параллельной DenseNet. Гибридная сверточная нейронная сеть с параллельными последовательными связями между слоями обладает более высокой эффективностью классификации тонов текста в 6 различные базы данных по сравнению с базовыми моделями CNN, TextCNN, FastText, DPCNN.

В ходе сравнительного анализа традиционных методов машинного обучения с сервисами Google с помощью Translate API A Shalkarbayuli, A Kairbekov and Y Amangeldi [7] выявили, насколько хорошо они работают с русскими текстами на традиционных методах машинного обучения после применения методов предварительной обработки, которые включали выделение, синтаксический анализ и индексацию, с использованием различных методов встраивания конвертировали неструктурированные данные в структурированные.

В статье Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments [8] авторы Nafis Irtiza Tripto и Mohammed Eunus Ali исследуют извлечение эмоций и анализ настроений в контексте бенгальского языка. Была создана модель, основанная на глубоком обучении для классификации предложений на языке бангла по трем классам (положительный, отрицательный, нейтральный) и пяти классам (сильно положительный, позитивный, нейтральный, отрицательный, сильно отрицательный) ярлыков настроения. Также была создана модель для извлечения эмоции из предложений в виде одной из шести основных эмоций (гнев, отвращение, страх, радость, печаль и удивление). Была произведена оценка производительности модели с использованием набора данных комментариев из различных видеороликов YouTube. Предложенный подход показал точность 65,97% и 54,24% для настроений трех и пяти меток соответственно.

Также авторами Darya Bogoradnikova, Anton Matveev и Olesia Makhnytkina [9] был предложен подход к анализу настроений и идентификации эмоций в комментариях пользователей. Решение состоит из трех частей: 1) определение темы, 2) оценка настроений, 3) обнаружение токсичности текста и локализация токсичных промежутков. Недостаток значительно больших обучающих данных для русского языка решается с помощью многоязычных встраиваний слов, модели адаптации к состязательной предметной области и увеличения данных. Представляется обзор различных конвейеров предварительной обработки для тематического моделирования и выделяется модель LDA-Mallet, которая демонстрирует наилучшую производительность. Для анализа настроений и обнаружения токсичности использовался метод опорных векторов и глубокая нейронная сеть с многоязычной языковой моделью и адаптацией к состязательной предметной области, которая позволяет обучать алгоритмы с наборами данных на английском языке. Все методы протестированы с использованием набора пользовательских комментариев к различным онлайн-курсам и скорректированы для обеспечения поддержки разработки виртуального диалогового помощника для проведения виртуальных экзаменов.

В статье Large Scale Legal Text Classification Using Transformer Models [10] авторами Zein Shaheen, Gerhard Wohlgenannt и Erwin Filtz решается проблема классификации большого текста с несколькими метками в правовой сфере, где наборы данных, такие как JRC-Acquis и EURLEX57K, помеченные словарем EuroVoc, были созданы в рамках правовых информационных систем Европейского союза. В этой работе изучается производительность различных моделей на основе трансформаторов в сочетании со стратегиями, такими как генеративная предварительная тренировка, постепенное размораживание и дискриминационные скорости обучения, чтобы достичь конкурентоспособных результатов классификации и представить новые современные результаты F1-метрика = 0,661 для JRC-Acquis и 0,754 для EURLEX57K. Также изучалась производительность языковых моделей трансформеров BERT, RoBERTa, DistilBERT и XLNet в заданных задачах классификации текста. BERT — это ранняя и очень популярная модель, RoBERTa - модифицированная версия BERT, обученная на большем корпусе, DistilBERT — это дистиллированная версия BERT и, следовательно, с меньшими вычислительными затратами, и, наконец, XLNet можно использовать с большими последовательностями входных токенов. Авторы используют предварительно обученные модели сетей. Входными данными для BERT является либо отдельный текст (предложение или документ), либо пара текстов. Первым маркером каждой последовательности является специальный классификационный маркер [CLS], за которым следуют маркеры слов для первого текста A, затем маркер-разделитель [SEP], и (необязательно) после этого лексемы WordPiece для второго текста B. В дополнение к вложениям токенов, BERT использует позиционные вложения для представления положения токенов в последовательности. Для обучения БЕРТ применяет задачи моделирования маскированного языка (MLM) и прогнозирования следующего предложения (NSP).

Pengfei Li и др. [11] представили внимательный сверточный преобразователь (ACT), который использует преимущества как Transformer, так и CNN для эффективной классификации текста. Был предложен новый механизм внимательной свертки, который использует семантическое значение сверточных фильтров для преобразования текста из сложного пространства слов в более информативное пространство сверточных фильтров, в котором фиксируются важные n-граммы. ACT способен эффективно фиксировать как локальные, так и глобальные зависимости, сохраняя при этом последовательную информацию. Эксперименты с различными задачами классификации текста и детальный анализ показывают, что ACT — это легкий, быстрый и эффективный универсальный классификатор текста, превосходящий CNNs, RNNS и другие модели, включая трансформеры.

В качестве **методов исследования** используется анализ статьи различных авторов и сравнение подходов. Для выявления частных проблем существующих исследований и усовершенствования процесса обработки текста используется метод дедукции – переход от общих концепций к созданию конкретной архитектуры нейронной сети. В процессе разработки трансформера и подбора гиперпараметров применялись экспериментальные методы. Для классификации отзывов использовались статистические методы.

## Заключение

Было выявлено, что область анализа тональности текста с использованием трансформеров является слабо исследованной на данный момент (особенно в российском сегменте из-за нехватки наборов данных) и актуальной ввиду современных тенденций развития российского рынка, т. к. все больше людей предпочитают приобретать товары и услуги онлайн, предварительно ознакомившись с отзывами. Многие отзывы могут иметь различную степень полезности для потенциального покупателя и их необходимо ранжировать по количеству ценной информации. Также анализ настроений может быть использован самими компаниями, выпускающими продукцию для улучшения качества товаров. Еще одной областью применения данной технологии являются рекомендательные системы, которые будут собирать информацию об отношении пользователя к различным аспектам товаров (косметической продукции, фильмов, игр и т. д.) и предлагать таргетированные варианты.

# СПИСОК ЛИТЕРАТУРЫ

[1] N. Archak, A. Ghose, P.G. Ipeirotis. Deriving the pricing power of product features by mining consumer reviews // Management Science. – Vol. 57 (8) - P. 1485–1509. – 2011.

[2] Jones, Q., Ravid, G., & Rafaeli, S. Information Overload and the Message Dynamics of Online Interaction Spaces: A Theoretical Model and Empirical Exploration // Information Systems Research. – Vol. 15(2), P. 194–210. – 2004. - doi:10.1287/isre.1040.0023

[3] Salnikova S., Kyrychenko R. Sentiment Analysis Based on the BERT Model: Attitudes Towards Politicians Using Media Data // Proceedings of the International Conference on Social Science, Psychology and Legal Regulation. – Vol. 617. – 2021.

[4] Vaswani A. et al. Attention Is All You Need // 31st Conference on Neural Information Processing Systems. – 2017.

[5] Xiaolin Li, Chaojiang Wu и Feng Mai. The Effect of Online Reviews on Product Sales: A Joint Sentiment-Topic Analysis // The International Journal of Information Systems Theories and Applications. – Vol. 56. – P. 172 -184. – 2018.

[6] Peleshchak R. et al. Text Tonality Classification Using a Hybrid Convolutional Neural Network with Parallel and Sequential Connections Between Layers // 6th International Conference on Computational Linguistics and Intelligent Systems. – 2022.

[7] Shalkarbayuli A., Kairbekov A. Amangeldi Y. Comparison of traditional machine learning methods and Google services in identifying tonality on Russian texts // Journal of Physics: Conference Series. Vol. 1117. – 2018. –doi:10.1088/1742-6596/1117/1/012002.

[8] Nafis Irtiza Tripto Mohammed Eunus Ali. Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments // 2018 International Conference on Bangla Speech and Language Processing (ICBSLP). – 2018. – DOI:10.1109/ICBSLP.2018.8554875

[9] Bogoradnikova D., Matveev A., Makhnytkina O. Multilingual Sentiment Analysis and Toxicity Detection for Text Messages in Russian // FRUCT. – 2021. – DOI: 10.23919/FRUCT52173.2021.9435584

[10] Shaheen Z., Wohlgenannt G., Filtz E. Large Scale Legal Text Classification Using Transformer Models. – 2020.

[11] Pengfei Li et. al. ACT: an Attentive Convolutional Transformer for Efficient Text Classification // The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21). – Vol. 35. – P. 13261-13269. – 2021. – DOI:10.1609/aaai.v35i15.17566