УДК 004.89

Разработка интеллектуальной системы аналитики текстовой информации с помощью больших языковых моделей

Я. И. Райко, О. А. Гудаев
ФГБОУ ВО «Донецкий национальный технический университет» (г. Донецк)
кафедра «Прикладная математика и искусственный интеллект»
E-mail: mnogorf@gmail.com

Аннотация
В работе рассматривается подход к созданию интеллектуальной системы аналитики текстовой информации в маркетинге на основе больших языковых моделей. Предложена схема бизнес-процессов, вариантов использования и архитектура системы. Внедрение системы, основанной на разработанных бизнес-процессах и архитектурных принципах, создаёт предпосылки для значительного повышения эффективности работы маркетинговых, продуктовых и аналитических подразделений, а также способствует ускорению принятия решений, повышению качества клиентской аналитики, своевременному выявлению трендов и рисков, а также формированию более точных и обоснованных стратегических рекомендаций для развития цифровых продуктов и коммуникаций с пользователем.

Введение

В маркетинговом анализе большая языковая модель (БЯМ) используется как универсальный семантический аналитик, который умеет понимать смысл текстов, выявлять намерения аудитории, сегментировать клиентов и автоматически формировать выводы, которые раньше требовали команды аналитиков. Компьютерная система на основе БЯМ используется как универсальный семантический интерпретатор, который позволяет работать с большими массивами текстовой информации не на уровне ключевых слов, а на уровне смыслов, намерений и эмоциональных оттенков. Модель читает отзывы, комментарии, обращения клиентов, сообщения в соцсетях. Затем автоматически выделяет темы, проблемы, мотивы поведения и скрытые паттерны, которые традиционные методы анализа не способны уловить.

Благодаря контекстному пониманию БЯМ выявляет эмоциональную окраску сообщений, различает иронию и сарказм, определяет истинные причины недовольства или наоборот – факторы лояльности. Она формирует динамические сегменты аудитории, группируя пользователей по смысловым признакам, а не по заранее заданным категориям, что позволяет маркетологам точнее понимать потребности разных групп. На основе анализа БЯМ генерирует аналитические выводы, краткие резюме больших массивов данных, формулирует гипотезы и рекомендации, фактически выполняя роль интеллектуального аналитика, который способен обработать тысячи текстов и представить их в виде структурированных гипотез-озарений.

Целью данной научной работы является разработка интеллектуальной системы аналитики текстовой информации, основанной на использовании больших языковых моделей, способной автоматически интерпретировать, классифицировать и обобщать неструктурированные текстовые данные для повышения качества аналитических выводов и поддержки принятия управленческих решений.

Основные методы разработки интеллектуальной системы аналитики текстовой информации на основе больших языковых моделей включают сочетание классических подходов обработки естественного языка и современных трансформерных архитектур. Применяется метод предобработки текстов, включающий нормализацию, токенизацию и выделение ключевых структур, что обеспечивает корректную подачу данных в модель. Для структурирования смысловых единиц используются методы извлечения сущностей и отношений, а также тематическое моделирование, позволяющее выявлять скрытые шаблоны в больших массивах текстов. Генерация аналитических выводов опирается на методы абдукции, как индукции, так и дедукции, что позволяет формировать краткие и содержательные отчёты.

Постановка задачи

Основная задача заключается в разработке интеллектуальной системы аналитики текстовой информации, способной автоматически интерпретировать, классифицировать и обобщать неструктурированные данные, возникающие в маркетинговой деятельности организаций.

Современные компании сталкиваются с постоянно растущими объёмами текстовых сообщений: отзывов, комментариев, обращений клиентов, публикаций в социальных сетях и внутренних документов. Традиционные методы анализа не обеспечивают достаточной глубины понимания контекста, эмоциональной окраски и скрытых смысловых шаблонов. В этой связи возникает необходимость создания системы, основанной на больших языковых моделях, которые обладают способностью к контекстному пониманию, генерации обобщений и извлечению значимых признаков из текстов.

Задача исследования состоит в формализации требований к такой системе, определении методов обработки данных, выборе архитектурных решений, обеспечивающих интеграцию языковой модели с корпоративной информационной средой, и разработке схемы бизнес-процессов и вариантов использования, демонстрирующих практическую применимость подхода.

Итогом решения задачи должно стать создание архитектуры и модели бизнес-процессов интеллектуальной системы, способных повысить качество маркетинговой аналитики и обеспечить поддержку принятия управленческих решений на основе глубокого анализа текстовой информации.

Существующие аналоги системы

Существующие аналоги интеллектуальных систем аналитики текстовой информации представлены как коммерческими платформами, так и исследовательскими решениями, использующими большие языковые модели для обработки неструктурированных данных. Наиболее распространённые системы ориентированы на анализ клиентских отзывов, мониторинг упоминаний фирмы и автоматическую классификацию обращений.

Существующие аналоги интеллектуальных систем аналитики текстовой информации в Российской Федерации представлены рядом коммерческих и корпоративных решений, ориентированных на обработку клиентских обращений, мониторинг репутации и анализ коммуникаций в цифровых каналах. Наиболее распространённые платформы, такие как Brand Analytics, YouScan и IQBuzz, используют методы машинного обучения и элементы современных языковых моделей для анализа социальных медиа, определения тональности сообщений и выявления ключевых тем обсуждений [1, 2, 3]. Эти системы широко применяются в маркетинге, однако их функциональность ограничена рамками предопределённых сценариев и не предполагает глубокой адаптации под внутренние бизнес-процессы конкретной организации.

В корпоративном секторе используются решения класса Contact Center AI, включая продукты Naumen, CleverDATA и Just AI, которые обеспечивают автоматическую классификацию обращений, маршрутизацию запросов и базовый смысловой анализ, но в большинстве случаев опираются на гибридные модели и не используют потенциал больших языковых моделей в полном объёме [2, 4, 5, 6].

В научно-исследовательской среде представлены прототипы систем, основанных на отечественных трансформерных моделях, включая семейства ruBERT, ruGPT и модели, разработанные в экосистеме Sber AI, однако эти решения чаще всего ориентированы на экспериментальные задачи и не обладают развитой архитектурой для интеграции с маркетинговыми процессами [7, 8, 9].

Крупные технологические компании предлагают универсальные облачные сервисы, такие как Google Cloud Natural Language, Amazon Comprehend и Microsoft Azure Text Analytics, которые обеспечивают базовые функции извлечения сущностей, определения тональности и тематической категоризации, однако их возможности ограничены отсутствием глубокой адаптации к предметной области маркетинга и невозможностью гибкой интеграции с корпоративными бизнес-процессами [10, 11, 12].

На рынке также присутствуют специализированные маркетинговые платформы, включая Sprinklr, Brandwatch и Clarabridge, использующие методы машинного обучения и элементы трансформерных моделей для анализа социальных медиа и клиентских коммуникаций. Они функционируют как закрытые экосистемы и не позволяют строить расширяемые архитектуры на основе собственных данных организации [1, 2, 3, 13].

Таким образом, несмотря на наличие отдельных инструментов для анализа текстовых данных, российские аналоги либо ограничены функционально, либо не обеспечивают комплексного подхода к интерпретации, классификации и обобщению маркетинговой информации, что подтверждает актуальность разработки специализированной интеллектуальной системы на основе больших языковых моделей.

Схема вариантов использования

Диаграмма вариантов использования разбита на две части. Первая часть показывает работу с первичными данными (см. рис.1).

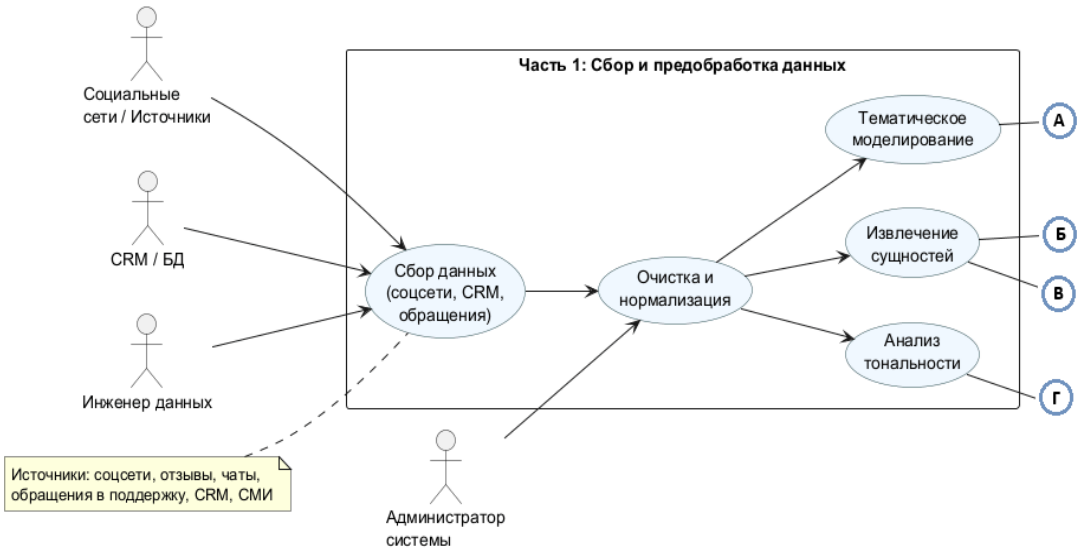

Рисунок 1 – Первая часть диаграммы вариантов использования

Диаграмма вариантов использования отражает целостную логику функционирования интеллектуальной системы аналитики текстовой информации как последовательный и взаимосвязанный процесс преобразования неструктурированных данных в управленческие решения. В её основе лежит структура, разбитая на две подсистемы, где первая часть описывает этапы сбора, очистки и нормализации данных, а вторая описывает механизмы аналитической обработки и доставку результатов конечным пользователям. На вход системы поступают сообщения из социальных сетей, CRM-систем и других внешних источников, что задаёт исходный поток данных, требующий унификации и подготовки. Инженер данных обеспечивает корректность подключения источников и стабильность каналов передачи, а администратор системы контролирует параметры обработки и конфигурацию модулей. На этом уровне система выполняет последовательные операции: агрегирует данные, устраняет шум и дубликаты, нормализует текст, выделяет сущности, определяет тональность и выявляет тематические структуры. Для этих операций подходит БЯМ. Операции образуют непрерывную цепочку <А, Б, В, Г>, где каждый шаг будет опираться на результаты предыдущего, формируя структурированное представление входных данных, пригодных для дальнейшего анализа.

Вторая часть диаграммы (рис. 2) описывает аналитический контур, в котором система преобразует извлечённые признаки в осмысленные выводы.

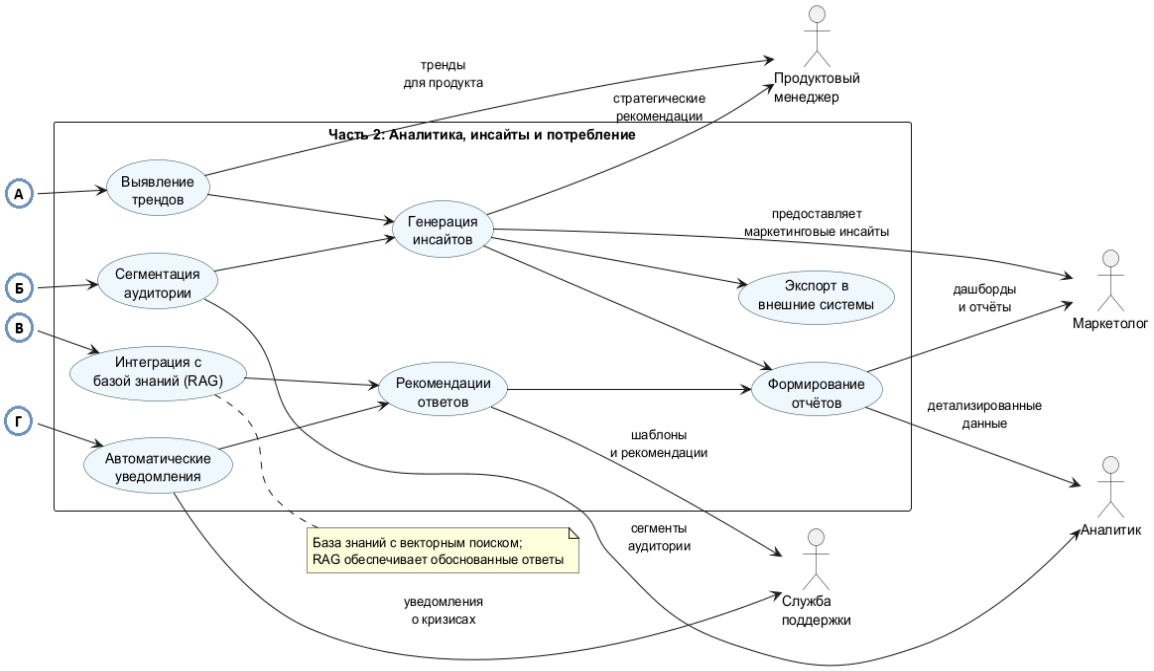

Рисунок 2 – Вторая часть диаграммы вариантов использования

Входная цепочка <А, Б, В, Г> получат данные от предыдущей подсистемы. Тематические модели и сущности служат основой для выявления трендов, а результаты тонального анализа позволяют фиксировать эмоциональные изменения и кризисные всплески. На основе этих данных формируются сегменты аудитории, выявляются закономерности поведения и создаются интерпретируемые инсайты, которые затем преобразуются в отчёты, визуализации и рекомендации. Модуль интеграции с базой знаний обеспечивает возможность обоснованной генерации ответов и уточнений, а система уведомлений автоматически информирует о значимых изменениях в информационном поле. Эти аналитические результаты направляются различным категориям пользователей: маркетолог получает стратегические предсказания и отчёты для планирования кампаний, продуктовый менеджер – данные о трендах и пользовательских ожиданиях, аналитик – детализированные показатели и сегментацию рынка, а служба поддержки – рекомендации по ответам и оперативные флажки внимания. Таким образом, диаграмма демонстрирует не набор разрозненных функций, а единую сквозную архитектуру, в которой каждый прецедент логически вытекает из предыдущего, а вся система работает как интегрированный механизм преобразования данных в знания, ориентированный на БЯМ поддержку принятия решений в маркетинге.

Моделирование бизнес-процессов системы

Для моделирования бизнес-процессов использованы UML-диаграммы деятельности, которые по структуре и логике очень близки к моделированию бизнес-процессов полноценным BPMN 2.0 (см. рис. 3).

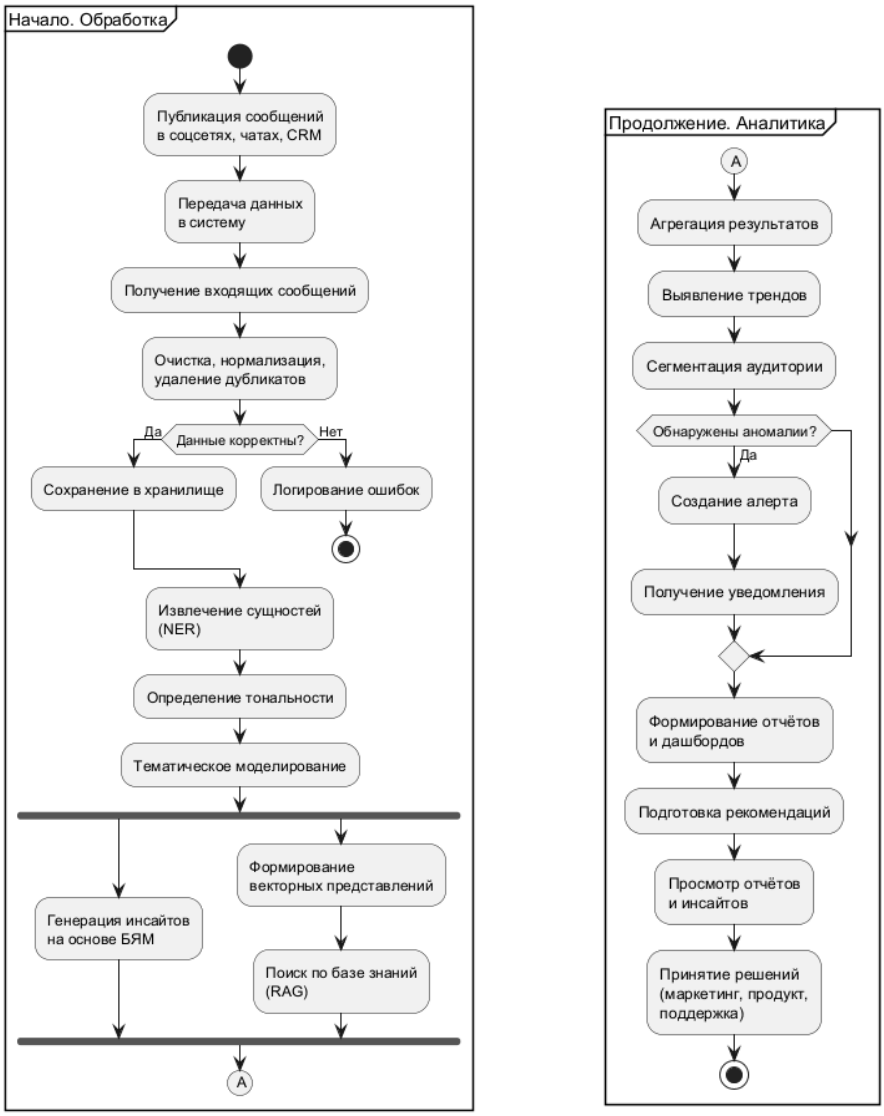

Рисунок 3 – Диаграмма активности

Бизнес-процессы системы аналитики, использующей большие языковые модели, представляют собой последовательный и взаимосвязанный цикл преобразования неструктурированных данных в аналитические выводы и управленческие решения. После подготовки данные передаются в модуль БЯМ, который является центральным вычислительным ядром системы. Здесь выполняется извлечение сущностей, определение тональности, тематическое моделирование и формирование векторных представлений. Эти операции позволяют преобразовать текст в структурированные признаки, пригодные для дальнейшего анализа. Параллельно модуль векторизации создаёт эмбеддинги и индексирует их, обеспечивая быстрый поиск релевантных фрагментов в базе знаний. Механизм RAG использует этот индекс для получения контекстной информации и формирования обоснованных ответов.

Следующий этап – аналитическая обработка, где агрегируются результаты и строятся временные ряды, выявляются тренды и сегменты аудитории. Этот процесс обеспечивает переход от локальных характеристик отдельных сообщений к обобщённым закономерностям, отражающим динамику информационного поля и поведение пользователей. Аналитический модуль формирует структурированные данные, которые служат основой для отчётности и визуализации. На заключительном этапе моделирования бизнес-процессов модуль отчётности преобразует аналитические результаты в дашборды, графики и текстовые рекомендации.

Архитектура системы

На рис. 4 представлена архитектура системы. В верхнем уровне архитектуры расположены источники данных, представленные внешними каналам. Следующим звеном является ETL-модуль, который выполняет очистку, нормализацию, устранение дубликатов и приведение данных к единому формату. Результаты этой обработки помещаются в хранилище подготовленных данных, которое служит центральным узлом для всех последующих этапов анализа. На основе подготовленных данных работает модуль БЯМ, включающий несколько специализированных компонентов.

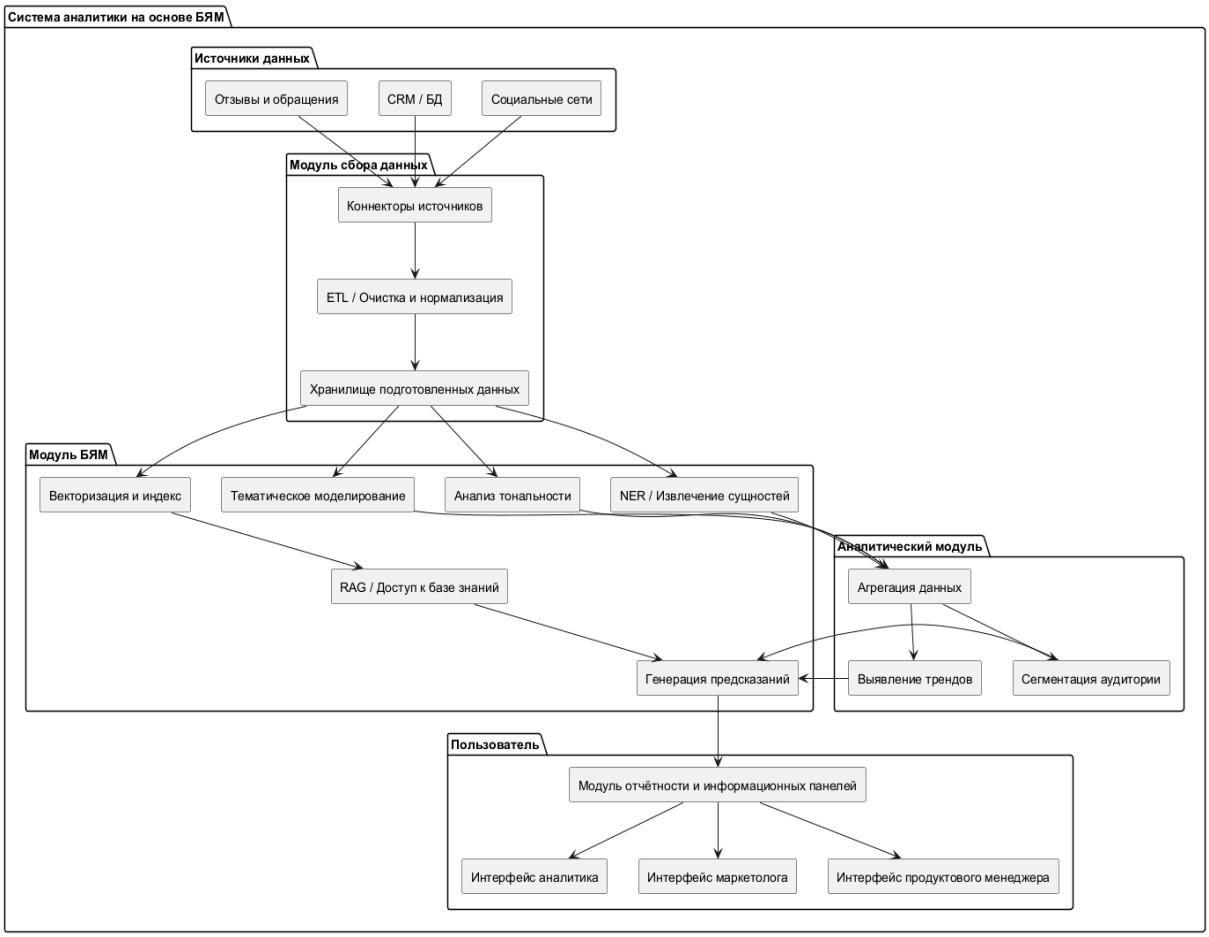

Рисунок 4 – Диаграмма компонентов архитектуры системы

Модуль извлечения сущностей выделяет ключевые объекты и атрибуты, модуль анализа тональности определяет эмоциональную окраску сообщений, а тематическое моделирование выявляет скрытые темы и кластеры обсуждений. Параллельно функционирует компонент векторизации, который преобразует тексты в эмбеддинги и формирует векторный индекс. Этот индекс используется модулем RAG для поиска релевантных фрагментов в базе знаний и последующей генерации обоснованных ответов. Завершающим элементом блока БЯМ является модуль генерации предсказания, который объединяет результаты всех предыдущих анализов и формирует интерпретируемые выводы.

Далее данные поступают в аналитический модуль, который агрегирует результаты, выявляет тренды, строит сегменты аудитории и формирует аналитические структуры, пригодные для визуализации и принятия решений. Финальный уровень архитектуры – пользовательский контур, включающий модуль отчётности и интерфейсы различных категорий пользователей: аналитиков, маркетологов и продуктовых менеджеров. Модуль отчётности формирует информационные панели, визуализации и рекомендации, которые затем отображаются в соответствующих пользовательских интерфейсах. Каждый интерфейс адаптирован под задачи конкретной роли: аналитик получает детализированные данные, маркетолог – стратегические предсказания, продуктовый менеджер – информацию о трендах и пользовательских ожиданиях.

Выводы

Проведённое исследование позволило сформировать целостное представление о том, как архитектурные и функциональные элементы системы аналитики маркетинга на основе больших языковых моделей объединяются в единый сквозной бизнес-процесс, обеспечивающий преобразование неструктурированных текстовых данных в интерпретируемые аналитические выводы.

Анализ жизненного цикла данных маркетинга показал, что ключевым фактором эффективности является согласованность этапов – от поступления информации и её нормализации до генерации предсказаний, выявления трендов и последующей визуализации результатов для конечных пользователей.

Особое значение имеет интеграция механизмов векторизации и RAG, которые обеспечивают обоснованность выводов и повышают качество интерпретации данных за счёт доступа к корпоративной базе знаний. Таким образом, исследование подтвердило, что применение БЯМ позволяет существенно расширить аналитические возможности системы, обеспечивая глубину понимания контекста, автоматизацию интерпретации и повышение точности аналитических моделей.

Внедрение системы, основанной на разработанных бизнес-процессах и архитектурных принципах, создаёт предпосылки для значительного повышения эффективности работы маркетинговых, продуктовых и аналитических подразделений. Централизованный конвейер обработки данных БЯМ-модулем обеспечивает воспроизводимость и прозрачность аналитики, а модульная архитектура позволяет масштабировать отдельные компоненты в зависимости от нагрузки и требований маркетинга. Использование БЯМ в сочетании с векторным поиском и механизмами RAG обеспечивает высокую адаптивность системы к новым источникам данных.

Внедрение такой системы способствует ускорению принятия решений, повышению качества клиентской аналитики, своевременному выявлению трендов и рисков, а также формированию более точных и обоснованных стратегических рекомендаций для развития цифровых продуктов и коммуникаций с пользователем.

Литература

1. Иванов, А. А. Анализ пользовательского контента в социальных сетях для оценки репутации бренда / А. А. Иванов, М. С. Лебедев // Маркетинг и маркетинговые исследования. – 2021. – № 4. – С. 22–31.

2. Соловьёв, И. В. Методы мониторинга социальных медиа в задачах управления клиентским опытом / И. В. Соловьёв, Е. А. Громова // Управление. – 2022. – № 3. – С. 45–54.

3. Киселёв, П. Н. Применение интеллектуальных систем для анализа отзывов и обращений потребителей / П. Н. Киселёв, Л. В. Сафонова // Информационные системы и технологии. – 2021. – № 6. – С. 112–120.

4. Морозова, Т. В. Технологии анализа больших данных в маркетинге: инструменты и подходы / Т. В. Морозова, А. П. Данилов // Вестник РГГУ. Серия: Экономика. – 2018. – № 4. – С. 89–98.

5. Махтибеков, А. Управление репутацией бренда через социальные медиа как стратегический инструмент PR / А. Махтибеков // Theoretical & Applied Science. – 2025. – № 4(144). – С. 24-28.

6. Федоров, А. М. Метод интеграции больших языковых моделей в алгоритмы фокусированного мониторинга открытых данных социальных медиа / А. М. Федоров, И. О. Датьев,

И. Г. Вишняков // Информатика и автоматизация. – 2025. – Т. 24, № 6. – С. 1623-1648.

7. Пимешков, В. К. Комбинированный метод извлечения терминов для задачи мониторинга тематических обсуждений в социальных медиа / В. К. Пимешков, М. Л. Никонорова, М. Г. Шишаев // Информатика и автоматизация. – 2024. – Т. 23, № 4. – С. 1110-1138.

8. Буравлев, А. С. Анализ качества реконструкции бизнес-процессов с помощью языковой модели ChatGPT / А. С. Буравлев, Д. Е. Демидова, Е. А. Ткачева // Техника средств связи. – 2025. – № 1(169). – С. 84-97.

9. Серова, В. С. Гибридный метод классификации текстовых данных с узкоспециализированной терминологией / В. С. Серова, А. В. Голлай, Е. В. Бунова // Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. – 2025. – Т. 25, № 3. – С. 42-52.

10. Мухамедиев, Р. И. Облачные сервисы для обработки текстов на естественном языке / Р. И. Мухамедиев, Л. А. Анварова, Я. Кучин, А. Сымагулов // Современные информационные технологии и ИТ-образование. – 2018. – Т. 14, № 3. – С. 540–551.

11. Технология обработки естественного языка / Microsoft Azure Architecture Center // Microsoft Learn. – 2023. – URL: https://learn.microsoft.com

12. Поспелов, Д. А. Применение методов машинного обучения для анализа текстовой информации в корпоративных системах / Д. А. Поспелов, Е. В. Крылова // Информационные системы и технологии. – 2021. – № 6. – С. 118–126.

13. Чернышов, А. В. Применение технологий искусственного интеллекта для автоматизации обработки текстовых данных / А. В. Чернышов, Л. С. Громова // Информационные технологии и телекоммуникации. – 2022. – Т. 10, № 1. – С. 15–27.

***Райко Я. И., Гудаев О. А. Разработка интеллектуальной системы аналитики текстовой информации с помощью больших языковых моделей**. В работе рассматривается подход к созданию интеллектуальной системы аналитики текстовой информации в маркетинге на основе больших языковых моделей. Предложена схема бизнес-процессов, вариантов использования и архитектура системы. Внедрение системы, основанной на разработанных бизнес-процессах и архитектурных принципах, создаёт предпосылки для значительного повышения эффективности работы маркетинговых, продуктовых и аналитических подразделений, а также способствует ускорению принятия решений, повышению качества клиентской аналитики, своевременному выявлению трендов и рисков, а также формированию более точных и обоснованных стратегических рекомендаций для развития цифровых продуктов и коммуникаций с пользователем.*

***Ключевые слова:** большие языковые модели, аналитика текста, маркетинг, интерпретация данных, классификация, бизнес-процессы, интеллектуальные системы.*

***Rajko Ya. I., Gudaev O. A. Development of an intelligent text analytics system using large language models.** This paper examines an approach to creating an intelligent text analytics system for marketing based on large language models. A diagram of business processes, use cases, and the system architecture are proposed. Implementation of a system based on developed business processes and architectural principles creates prerequisites for a significant increase in the efficiency of marketing, product, and analytical departments, as well as contributes to faster decision-making, improved quality of customer analytics, timely identification of trends and risks, and the formation of more accurate and well-founded strategic recommendations for the development of digital products and user communications.*

***Keywords:** large language models, text analytics, marketing, data interpretation, classification, business processes, intelligent systems.*

Статья поступила в редакцию 18.11.2025
Рекомендована к публикации профессором Павлышом В. Н.