DOI 10.37882/2223-2966.2025.02-2.34

АНАЛИЗ И ОПТИМИЗАЦИЯ ПРОЦЕССОВ В ТЕХНИЧЕСКИХ СИСТЕМАХ С ПРИМЕНЕНИЕМ ОБРАБОТКИ ТЕКСТОВЫХ ДАННЫХ

ANALYSIS AND OPTIMIZATION OF PROCESSES IN TECHNICAL SYSTEMS USING TEXT DATA PROCESSING

A. Tereshenko

Summary. Modern technical systems combine hardware, software, and information resources to form complex structures that critical functions depend on. A significant portion of the data in such systems is presented in text form, including technical documentation, reports, and protocols. However, existing methods for processing text information are not sufficiently adapted to work with technical vocabulary and specialized contexts, which limits their use in control, diagnostics, and forecasting tasks. As part of the study, methods for processing text data were developed aimed at optimizing processes in technical systems. The proposed approach is based on the integration of methods of system analysis, machine learning, and text processing. The main focus is on creating specialized models that take into account the specifics of technical information, structure and classify data, and predict their impact on key processes. The results of the work contribute to increasing the reliability, performance, and adaptability of technical systems. The study has practical and scientific significance, providing tool solutions for analyzing text data in engineering and laying the foundation for further developments in this area.

Keywords: technical systems, text data processing, machine learning, system analysis, technical documentation, text information, process optimization, forecasting, diagnostics, engineering.

Терешенко Андрей Алексеевич
Аспирант, ФГАОУ ВО «Северный (Арктический) Федеральный Университет имени М.В. Ломоносова», г. Архангельск
andrey.tereshenko2017@mail.ru

Аннотация. Современные технические системы объединяют оборудование, программное обеспечение и информационные ресурсы, образуя сложные структуры, от которых зависят критически важные функции. Существенная часть данных в таких системах представлена в текстовой форме, включая техническую документацию, отчеты и протоколы. Однако существующие методы обработки текстовой информации недостаточно адаптированы для работы с технической лексикой и специализированными контекстами, что ограничивает их применение в задачах управления, диагностики и прогнозирования. В рамках исследования разработаны методы обработки текстовых данных, ориентированные на оптимизацию процессов в технических системах. Предложенный подход основан на интеграции методов системного анализа, машинного обучения и обработки текстов. Основное внимание уделено созданию специализированных моделей, которые учитывают особенности технической информации, структурируют и классифицируют данные, а также позволяют прогнозировать их влияние на ключевые процессы. Результаты работы способствуют повышению надежности, производительности и адаптивности технических систем. Исследование имеет практическую и научную значимость, предоставляя инструментальные решения для анализа текстовых данных в инженерии и закладывая основу для дальнейших разработок в этой области.

Ключевые слова: технические системы, обработка текстовых данных, машинное обучение, системный анализ, техническая документация, текстовая информация, оптимизация процессов, прогнозирование, диагностика, инженерия.

На сегодняшний день текстовые данные представляют собой неструктурированную информацию, включающую документы, отчёты, инструкции и комментарии, играющие ключевую роль в проектировании, эксплуатации и обслуживании технических систем. Они содержат сведения о характеристиках оборудования, результатах анализа, рекомендациях и реальных условиях эксплуатации. В отличие от структурированных данных, текстовые данные требуют специализированных методов анализа, что делает их обработку сложной, но ценной для оптимизации и модернизации систем. Их универсальность и доступность обеспечивают широкое применение в информационных технологиях и системном проектировании. Текстовые данные в технических системах классифицируются по структуре, формату, источнику, назначению, сложности и степени структурированности. Такая классификация позволяет эффективно обрабатывать текстовую информацию для оптимизации работы технических систем. Обработка текстовых данных включает извлечение, анализ и представление информации с использованием методов предобработки, машинного обучения и текстовых редакторов. Предобработка очищает данные с помощью токенизации, нормализации, удаления стоп-слов, лемматизации и стемминга, упрощая дальнейший анализ. Методы анализа, такие как частотный и семантический анализ, помогают выявить ключевые темы, связи и эмоции текста. Машинное обучение автоматизирует задачи классификации, извлечения информации и создания виртуальных помощников, применяя RNN, CNN и трансформеры для эффективной обработки текста. Текстовые данные обеспечивают инженеров ключевой инфор-

мацией для разработки, проектирования и внедрения технологий. Они включают технические документы, научные статьи, аналитические отчеты и учебные материалы, способствуя изучению решений, сравнительному анализу технологий и генерации инновационных идей. Текстовые данные важны для командной коммуникации через документацию процессов и обратную связь, а также для повышения квалификации инженеров через книги, статьи и кейс-стадии. Их обработка позволяет выявлять закономерности, проблемы и новые возможности, что улучшает эффективность инженерного творчества и инноваций.

На сегодняшний день технический объект представляет собой основную часть разработанной системы, которая направлена на автоматизацию процессов обработки и анализа текстовых данных. Технический объект — это созданное человеком или автоматом устройство, предназначенное для удовлетворения потребностей, будь то отдельная машина, прибор, здание, технологическая линия или программное обеспечение. Технический объект включает элементы, такие как узлы и детали, а также комплексы машин и систем, например, заводы или цеха. Основная цель — удовлетворение потребностей человека при соблюдении целесообразности и эффективности. В данном исследовании в качестве технического объекта рассмотрим автоматизированную систему обработки текстовых данных, которую можно будет применить в науке, бизнесе и промышленности для анализа информации, а также оптимизации решений.

Проблема исследования заключается в том, что современные методы обработки текстовой информации недостаточно эффективны для работы с данными технических систем, где важны специфика терминологии и контексты. Для решения нужно разработать методы, учитывающие эти особенности, с использованием системного анализа, машинного обучения и обработки текстов. Проблемы включают сложность данных, высокие требования к точности и отсутствие универсальных алгоритмов. Решение улучшит точность и скорость обработки, что приведет к снижению затрат и повышению надежности систем, а также ускорит технологический прогресс и рост конкурентоспособности.

Существующие системы анализа текстов имеют несколько существенных проблем. Во-первых, они часто ориентированы на общий текст и плохо адаптируются под специфические задачи, что ограничивает их применение в таких областях, как промышленность или наука. Во-вторых, они не учитывают контекст, тональность и эмоциональную окраску текста, что важно при анализе данных, связанных с человеческими факторами, например, отзывов и комментариев. Это приводит к необходимости ручного анализа, который может быть субъективным и подверженным ошибкам. Для улучшения ситуации необходимы специализированные системы, учитывающие контекст и минимизирующие влияние человеческого фактора. Также важно снизить сложность настройки системы под новые области знаний и повысить ее адаптивность. Для улучшения систем целесообразно добавить новые функциональные элементы, например, модуль семантического анализа для учета контекста текстов, компонент оценки тональности для анализа отзывов и социальных медиа, а также механизм самообучения для адаптации к изменяющимся условиям. Удобный интерфейс позволит гибко настраивать систему под конкретные задачи. Кроме того, необходимо исключать устаревшие или избыточные элементы, такие как старые алгоритмы или функции, которые усложняют систему. Оптимизация достигается перераспределением функций, например, перенаправлением первичной фильтрации данных в компонент предварительной обработки. Разделение элементов, совмещающих несколько функций, также повысит эффективность, например, разделение универсального модуля анализа текста на модули для лексического, синтаксического и семантического анализа.

Разработка автоматизированной системы обработки текстовых данных должна обеспечивать высокую эффективность. Система должна обеспечивать точность извлечения ключевых данных не менее 95 % и распознавание специализированной технической терминологии на уровне 98 %. Она должна поддерживать обработку текстов разного объема, включая крупные массивы данных. Технологическая эффективность включает производительность, то есть большие объемы текста должны обрабатываться эффективно. Система должна поддерживать русский язык и учитывать технические термины и контекст. Экономическая целесообразность заключается в минимизации затрат на внедрение и эксплуатацию, а также оптимизации энергопотребления и вычислительных ресурсов. Для улучшения взаимодействия с пользователем интерфейс должен быть интуитивно понятным, с возможностью гибкой настройки параметров анализа и визуализацией результатов, что снижает когнитивную нагрузку и повышает производительность. Как итог, идеальное техническое решение для автоматизированной системы обработки текстовых данных заключается в создании платформы, обеспечивающей точное и быстрое извлечение информации из текстов любых форматов и объемов. Система должна быть автономной, обучаемой и адаптируемой к новым областям с минимальным вмешательством человека. Важным элементом является использование методов машинного обучения и обработки естественного языка (NLP), что позволяет учитывать контекст, специализированную лексику и работать с текстами на разных языках. Система должна интегрироваться с внешними источниками данных и базами знаний, иметь модульную архитектуру для масштабирования и добавления новых функций. Она

должна минимизировать участие человека в рутинных процессах, позволяя сосредоточиться на интерпретации результатов и принятии решений, что повысит производительность и точность, делая систему незаменимым инструментом в промышленности, науке и бизнесе.

Для достижения описанных выше аспектов было разработано несколько алгоритмов. Разработка алгоритмов для модели является ключевым элементом исследования, направленным на автоматизацию анализа больших объёмов текстовых данных. Первый разработанный алгоритм начинается с автоматического запуска процесса обучения при изменении данных. Он проверяет обновления в датасете, и, если изменения обнаружены, запускает повторное обучение модели. В процессе обучения выполняется предварительная обработка данных, включая создание словаря и векторизацию текстов. Затем обучается модель логистической регрессии, и обновлённые данные сохраняются для дальнейшего использования. Этот процесс обеспечивает актуализацию модели, что важно для поддержания высокой точности и эффективности. Также был разработан алгоритм для автоматической классификации текста по эмоциональной окраске. Он начинается с добавления текста пользователем, после чего текст проходит предварительную обработку: токенизацию, удаление стоп-слов и преобразование в векторное представление с использованием словаря. Далее текст анализируется для определения тональности и классификации. Алгоритм оценивает степень уверенности модели в классификации, распределяя баллы: если Score <0,49 — негативный, = 0,50 — нейтральный,> 0,51 — положительный. Это позволяет точно интерпретировать результаты, предоставляя количественную меру уверенности. Уникальность алгоритма заключается в адаптации для русскоязычных технических текстов, учитывающих их специфические особенности и терминологию, что делает модель особенно полезной в технических системах. В итоге, архитектура

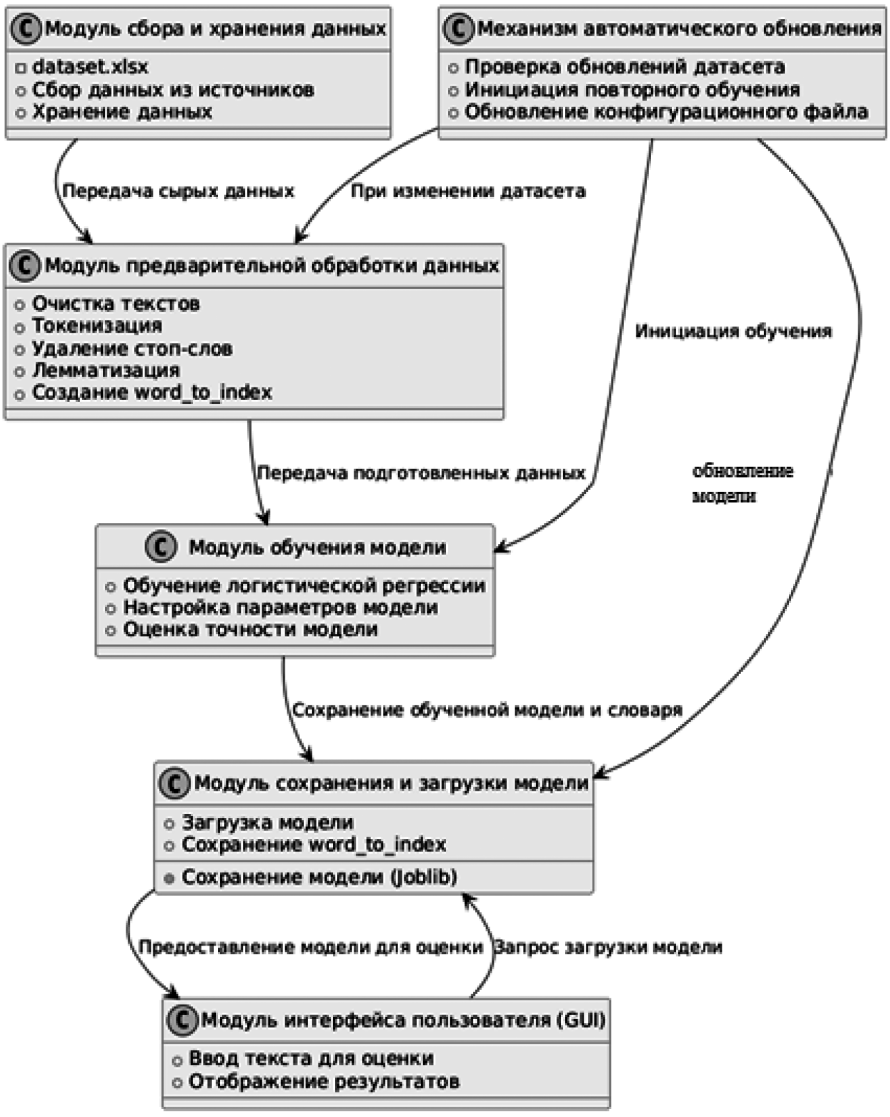

Рис. 1. Общая схема архитектуры модели

модели представляет собой интегрированную систему для анализа и классификации текстовых данных на русском языке, основываясь на модели логистической регрессии, обученной на специализированных данных, рисунок 1.

Система включает модули сбора данных, предварительной обработки (токенизация, очистка, удаление стоп-слов), обучения модели, сохранения и загрузки модели, а также интерфейс пользователя для ввода текстов и просмотра результатов. Ключевая особенность архитектуры — автоматическое обновление модели при изменении данных, что поддерживает актуальность и точность без необходимости ручного вмешательства, обеспечивая гибкость и адаптивность системы.

Далее было разработано приложение для удобства работы пользователя при анализе текстовых данных. Оно предоставляет интуитивно понятный интерфейс для ввода и обработки текстовых и аудиоданных, а так-

Рис. 2. Интерфейс и работа приложения

же визуализации результатов анализа, что помогает оптимизировать процессы в технических системах. Приложение построено с использованием Python, Tkinter для интерфейса и Joblib и Scikit-learn для работы с моделью. Модульная архитектура приложения обеспечивает гибкость, расширяемость и простоту использования. Система автоматически обновляет модель, что повышает её точность и эффективность, а также позволяет интегрировать новые функции в будущем. Для оптимизации взаимодействия пользователя с системой, особенно при работе с текстовыми данными, предусмотрены альтернативные способы ввода, такие как голосовое введение и загрузка аудиофайлов. Для реализации транскрипции речи интегрирован модуль VOSK, поддерживающий более 20 языков и диалектов, включая русский. Этот модуль позволяет эффективно распознавать речь в офлайн-режиме на устройствах с ограниченными ресурсами, имеет компактный размер (около 50 Мб) и поддерживает потоковый API с нулевой задержкой, а также динамическую настройку словаря и идентификацию говорящих. Как итог, для удобного взаимодействия пользователей с моделью анализа текстовой информации была разработана система упрощённого интерфейса. В приложении интегрирована разработанная ранее модель для анализа текста и транскрипции аудио в текст, что позволяет пользователям использовать функционал без глубоких знаний в области машинного обучения, рисунок 2.

Взаимодействие с моделью осуществляется через модуль, который обрабатывает текст (токенизация, удаление стоп-слов) и передает его на классификацию. Результаты возвращаются обратно в интерфейс. Также предусмотрен модуль автоматического обновления, который отслеживает изменения в датасете и обновляет модель. Это обеспечивает высокую точность и актуальность без вмешательства пользователя.

Итоговый вывод заключается в том, что данное исследование представляет собой значительный шаг в оптимизации процессов обработки и анализа текстовых данных, что напрямую влияет на повышение эффективности принятия решений в различных сферах, включая промышленность, бизнес, науку и управление. Внедрение автоматизированной системы, сочетающей методы обработки естественного языка (NLP) и машинного обучения, позволяет существенно ускорить и улучшить качество анализа больших объемов текстовой информации, что ранее требовало значительных временных и человеческих ресурсов. Эффективная обработка текстовых данных, включая их классификацию, предоставляет пользователю инструменты для более точного и обоснованного принятия решений в условиях неопределенности. Одним из глобальных достижений данной разработки является возможность динамического обновления модели, что способствует поддержанию её актуальности и точности в условиях постоянного изменения данных. Это позволяет системе адаптироваться к новым источникам информации и быстро реагировать на изменения в окружении, обеспечивая оперативное принятие решений. В традиционных системах анализа данные часто устаревают, что снижает их ценность для принятия обоснованных решений. Автоматизация процесса обновления модели исключает человеческий фактор в этом процессе, что минимизирует вероятность ошибок и повышает надёжность принимаемых решений. Кроме того, интеграция различных функциональных компонентов — от моделирования и обработки текстов до транскрипции аудио и визуализации данных — позволяет пользователю взаимодействовать с системой через интуитивно понятный интерфейс, что упрощает доступ к аналитической информации без необходимости глубоких знаний в области технологий машинного обучения. Это расширяет круг пользователей, которые могут эффективно работать с системой, что в свою очередь способствует более широкому применению в различных организационных контекстах, включая бизнес-аналитику, стратегическое планирование и управление рисками. Также значительно снижается время, необходимое для получения результатов, и повышает точность этих результатов. Это позволяет организациям быстрее реагировать на изменения в информации, выстраивать более обоснованные стратегии и принимать более точные решения. Таким образом, разработка не только улучшает процессы анализа и обработки текстовых данных, но и способствует более высокому уровню автоматизации принятия решений, что существенно повышает общую эффективность управления и снижает риски, связанные с недостаточной точностью анализа.

ЛИТЕРАТУРА

1. А.И. Половинкин. Основы инженерного творчества. Учебное пособие для вузов, 8-е изд., стер. ISBN: 978-5-507-45273-6, 362 стр, 2019 год.
2. Волкова В.Н., Денисов А.А. Теория систем и системный анализ / Изд. 3-е. М.: Юрайт, 2022. — 562 с.
3. Тарасенко, Ф.П. Прикладной системный анализ: учебное пособие / Ф. П. Тарасенко. — 2-е изд. — М.: КноРус, 2021. — 321 с.
4. Метрики в задачах машинного обучения [Электронный ресурс]. — URL: https://habr.com/ru/companies/ods/articles/328372/ (дата обращения: 26.10.2024).
5. Метрики в машинном обучении: понимание, применение и интерпретация [Электронный ресурс]. — URL: https://shakhbanov.org/metriki-v-mashinnom-obuchenii/ (дата обращения: 26.10.2024).

Журнал «Современная наука: актуальные проблемы теории и практики»