ISSN 1992-6502 (Print) Вестник УГАТУ ISSN 2225-2789 (Online)
2026. Т. 30, № 2 (112). С. 41-48 http://journal.uust.ru/index.php/Vestnik

УДК 004.06 doi 10.54708/19926502_2026_30211241
Код ГРНТИ 81.93.29

Разработка модуля автоматизированной предобработки текстовых данных с адаптивным управлением для нейросетевых моделей

Д.С. Алексеева[1*], Е.В. Юлина[2], Н.А. Кононов[1], В.В. Антонов[1], А.С. Дьячков[3]

[1]ФГБОУ ВО Уфимский университет науки и технологий, г. Уфа, Республика Башкортостан, Россия
[2]ООО «Центр Информационных Технологий», г. Уфа, Республика Башкортостан, Россия
[3]ООО «Газпромнефть-Цифровые решения», г. Санкт-Петербург, Россия

Аннотация. В статье рассматривается разработка модуля системы автоматизированной обработки данных для подготовки датасетов перед обучением нейронных сетей. Целью работы является преодоление проблем, связанных с низким качеством и неоднородностью входных данных, которые приводят к снижению производительности и переобучению моделей. В статье предложена комплексная архитектура модуля, включающая формальную декомпозицию процесса предобработки, многоуровневый контур управления для оперативного регулирования и адаптации, а также детализированную модель данных и схему базы данных на основе СУБД MariaDB. Особое внимание уделено методам обработки текстовой информации в задачах NLP. Представленное решение позволяет стандартизировать и значительно ускорить этап подготовки данных, минимизировав ручной труд и влияние человеческого фактора. Результаты работы будут полезны исследователям и разработчикам, стремящимся повысить качество, воспроизводимость и эффективность моделей машинного обучения за счет надежной автоматизации предварительной обработки данных.

Ключевые слова: автоматизированная обработка данных, предобработка данных, нейронные сети, машинное обучение, обработка естественного языка.

***ads.stat@mail.ru**

Введение

Нейронные сети, как известно, демонстрируют высокую эффективность при решении сложных задач в области машинного обучения. Однако качество их обучения напрямую определяется входными данными [1]. Реальные массивы данных нередко содержат пропуски, выбросы и дубликаты, а перечисленные недостатки, в свою очередь, ведут к падению производительности модели, ее переобучению, искажению результатов и росту временны́х затрат на обучение [2].

Чем шире становится сфера применения нейронных сетей, тем очевиднее ограничения ручной обработки данных [3]. Среди них – высокая трудоемкость, длительность процесса, неизбежный риск ошибок, обусловленный человеческим фактором, нестабильность итогового качества в зависимости от исполнителя, а также трудности масштабирования на большие объемы информации и потребность в узкопрофильных экспертных знаниях. Все это делает ручную обработку неэффективной и дорогостоящей, особенно когда речь идет о крупных датасетах, необходимых для обучения современных нейросетей.

Анализ текстовых данных сопряжен с несколькими ключевыми проблемами, которые были выявлены и систематизированы в работе [4].

Первая из них – неструктурированность и неоднородность текстов. Эти свойства проявляются через грамматические ошибки, опечатки, сленг и аббревиатуры, что серьезно затрудняет анализ: осложняется извлечение информации и понимание контекста, падает качество результатов. Как следствие, возникает необходимость в разработке алгоритмов нормализации текста.

Вторая проблема связана с высокой размерностью признакового пространства, которая неизбежна при векторном представлении текстов методами «Bag of Words» или TF-IDF. Это приводит к так называемому «проклятию размерности» и снижает эффективность обучения [5]. На практике указанное явление мешает обобщению и провоцирует переобучение: модель

перестает корректно работать с новыми данными. Поэтому требуются методы снижения размерности и отбора признаков.

Третья проблема – семантическая неоднозначность, когда одно и то же слово в разных контекстах приобретает различный смысл. Решить ее помогают методы разрешения неоднозначности, в частности контекстуальные векторные представления слов, учитывающие окружающее лексическое окружение [6, 7].

Отдельную сложность представляет языковая зависимость: методы предобработки часто специфичны для конкретного языка и требуют адаптации, учитывающей различия в порядке слов, грамматических конструкциях и морфологии. Для решения этой проблемы мы разрабатываем универсальные алгоритмы, способные адаптироваться к различным языковым системам, в русле тенденций, отраженных в мультиязычных моделях [8]. Наконец, существует проблема обработки редких слов, отсутствующих в словаре, для решения которой мы рассматриваем применение подсловных единиц [9] и алгоритмов работы с неизвестными токенами, что обеспечивает полноту обработки текстов независимо от встречаемости лексических единиц. В совокупности эти проблемы формируют комплексную задачу, для решения которой мы предлагаем интегрированный подход, сочетающий лингвистические методы с алгоритмами машинного обучения для создания эффективной системы предобработки текстовых данных.

В связи с этим эффективная обработка данных становится критическим этапом разработки модели машинного обучения, требующим специализированных инструментов. В данной статье мы предлагаем архитектуру модуля автоматизированной обработки данных для подготовки датасетов перед обучением нейронной сети.

Предлагаемый метод допускает интеграцию с компактными моделями. Последние особенно востребованы в условиях роста числа IoT-устройств и необходимости локальной обработки данных – это направление известно в литературе как TinyML [10]. Архитектура модуля, кроме того, позволяет подключать дополненную аналитику, что упрощает анализ данных для пользователей без глубоких технических знаний. Что касается предметной области, модуль нацелен на автоматизированную предобработку текстовых данных. В нее входят устранение шума, нормализация, структурирование и трансформация. Все перечисленные операции повышают эффективность нейросетевых моделей в задачах обработки естественного языка (NLP) [11].

Основной целью статьи является разработка методологии и архитектуры модуля для комплексной автоматизации подготовки данных, направленной на повышение качества входных датасетов, эффективности и воспроизводимости процесса обучения нейронных сетей. Для достижения этой цели решаются задачи:

− анализ ключевых проблем и рисков, связанных с качеством входных данных и ручной предобработкой в конвейере машинного обучения;

− разработка системы управления для обеспечения устойчивости и гибкости процесса обработки данных;

− определение структуры данных для хранения метаинформации о наборах данных, этапах обработки, моделях и пользователях системы;

− интеграция математических подходов для повышения качества формализации данных.

Научная новизна работы заключается в создании комплексной архитектуры, в которой математический метод интегрирован в сквозной управляемый процесс автоматизированной предобработки, обеспечивая единство формального описания, исполнения и контроля качества данных.

Современные подходы к обработке данных

Предобработка текстовых данных обычно строится вокруг нескольких стандартных этапов. Начинают с разбиения текста на слова, символы или подслова. Для этого применяются алгоритмы по типу whitespace tokenization или byte-pair encoding. Следующий шаг – удаление

стоп-слов, лексем без значимой смысловой нагрузки. Такая фильтрация позволяет модели сосредоточиться на информативных терминах.

Лемматизация и стемминг служат для приведения слов к нормальной либо корневой форме. Оба подхода снижают размерность данных и повышают их однородность. После этого текст векторизуют. Алгоритмы Bag of Words и TF-IDF подходят для базового частотного анализа, Word2Vec и GloVe учитывают семантические связи между словами, BERT же позволяет работать с контекстом на уровне целых предложений.

Пропуски в данных обрабатывают двумя основными способами: заполняют пропущенные значения, либо удаляют неполные записи. И то и другое повышает качество данных и надежность итоговых результатов.

Таким образом, грамотно выстроенная предобработка не только решает типичные проблемы текстового анализа, но и создает базу для выявления более сложных закономерностей. Сегодня, когда объемы данных растут экспоненциально, а их источники становятся все разнообразнее, эффективная предобработка превращается в критическое условие успешного применения машинного обучения – будь то в бизнесе или в науке.

Формальная формулировка проблемы обработки данных представлена на Рис. 1, а декомпозиция задачи иллюстрируется на Рис. 2.

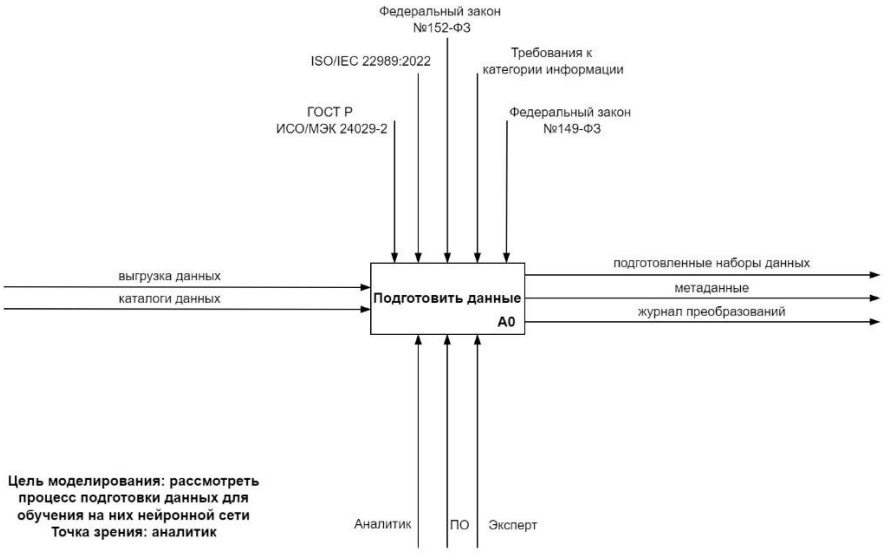

Рисунок 1. Контекстная диаграмма процесса подготовки данных.

На Рис. 2 изображена декомпозиция процесса подготовки данных. В основе модуля обработки данных лежит поэтапный подход. Сначала происходит проверка данных, включающая верификацию выгрузки данных и каталогов данных. Этот этап осуществляется с учетом требований к качеству контента, а также в соответствии с Федеральными законами № 152-ФЗ и № 149-ФЗ. Далее данные нормализуются, а затем разделяются на репрезентативные и нерепрезентативные выборки. На заключительном этапе производится форматирование данных в соответствии со стандартами ГОСТ Р ИСО/МЭК 24029-2 и ISO/IEC 22989:2022. Аналитик взаимодействует с модулем, отправляя запросы и получая результаты обработки.

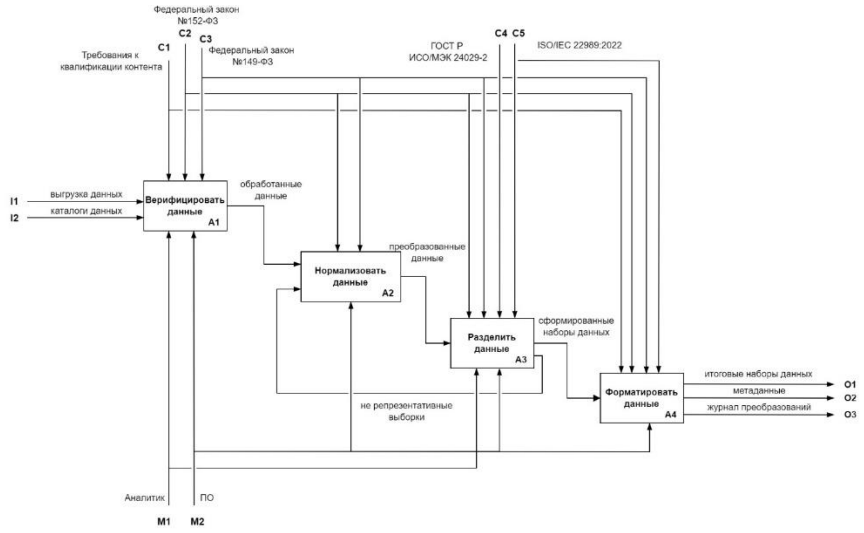

Рисунок 2. Декомпозиция процесса подготовки данных.

Представленная на Рис. 3 физическая модель данных включает шесть основных таблиц: «Аналитик», «Источник данных», «Набор данных», «Признак», «Модель машинного обучения» и «Этап обработки». Каждая таблица имеет свой набор атрибутов и связана с другими таблицами через идентифицирующие (сплошные линии) и неидентифицирующие (пунктирные линии) связи, что и определяет структуру хранения и взаимосвязи между данными. Таблица «Аналитик» хранит информацию о пользователях системы, а «Источник данных» – детали о местах хранения данных. Таблица «Набор данных» содержит метаданные о наборах данных, используемых для анализа, тогда как «Признак» описывает отдельные признаки этих наборов. В таблице «Модель машинного обучения» хранится информация о моделях, разработанных в системе, а в «Этапе обработки» – детали этапов обработки данных. Вместе эти таблицы образуют структуру, поддерживающую сбор, хранение, обработку и анализ данных в аналитической системе.

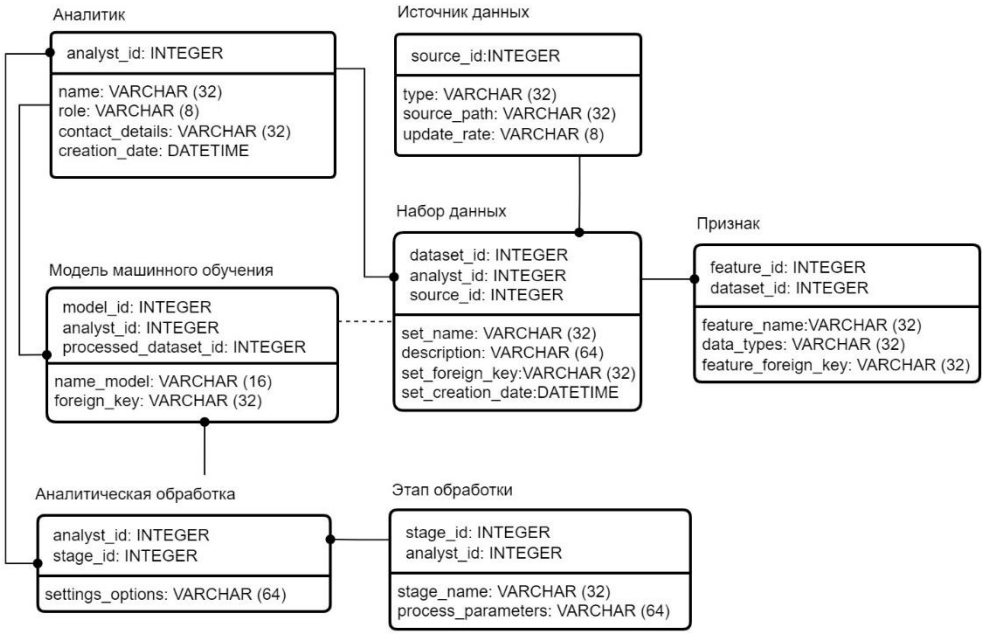

Рисунок 3. Физический уровень представления.

В основе архитектуры системы моделирования данных лежат шесть сущностей: «Аналитик», «Источник данных», «Набор данных», «Признак», «Этап обработки» и «Модель

машинного обучения». Эти сущности связаны в граф; целостность данных обеспечивается за счет первичных и внешних ключей.

Аналитик выступает как субъект системы. О нем хранятся следующие сведения: ФИО, роль (она же права доступа), контактные данные, метка времени регистрации и первичный ключ analyst_id. Создание наборов данных, настройка пайплайнов и запуск обучения моделей – все эти действия фиксируются в журнале операций в качестве исходных событий.

Источник данных определяет, откуда поступают необработанные данные. Его атрибуты: source_id (первичный ключ), тип источника, путь к данным и периодичность обновления.

Набор данных – это подготовленный к обработке снимок данных. Кроме технических атрибутов он содержит ссылки на родительский источник (source_id) и на ответственного аналитика (analyst_id). Благодаря этому фиксируются происхождение данных и правообладание.

Признак описывает отдельную переменную внутри набора данных. Его атрибуты: feature_id (первичный ключ), название признака, тип данных (числовой или категориальный) и внешний ключ dataset_id.

Этап обработки представляет собой одну операцию трансформации в пайплайне предобработки. Каждый этап идентифицируется по stage_id и задается двумя параметрами: алгоритмом (нормализация, one-hot encoding и т.п.) и вектором параметров. Последовательность этапов выстраивается в граф вычислений (DAG).

Модель машинного обучения – это результат выполнения пайплайна. Модель (model_id) обязательно привязывается к идентификатору обработанного набора данных (processed_dataset_id) и к аналитику (analyst_id). Такая привязка гарантирует полную воспроизводимость этапа формирования модели.

На основе описанной структуры реализуются четыре процесса.

Загрузка данных из внешних источников. Источник задает параметры подключения; загруженные данные связываются с набором через source_id.

Создание набора данных аналитиком. Аналитик отвечает за создание нового набора; связь фиксируется через analyst_id.

Предобработка данных. К любому набору данных можно применить цепочку этапов обработки. Аналитик настраивает эту цепочку через промежуточную таблицу AnalystProcessingStage, которая как раз и связывает аналитика с конкретным этапом.

*Обучение модели***.** Модель обучается на обработанном наборе данных – для этого служит внешний ключ processed_dataset_id. Аналитик, запустивший обучение, указывается через analyst_id.

Управление признаками. Признаки каждого набора данных описываются в отдельной таблице; связь признака с набором поддерживается через dataset_id.

Поскольку атрибуты набора данных, включая текстовые признаки, требуют преобразования в числовое представление для обработки моделями машинного обучения, ключевую роль играет этап векторизации. Для оценки качества этого преобразования и смысловой близости полученных векторных представлений используется математический аппарат. Одна из базовых формул (формула 1), рассчитывающая семантическую дистанцию между векторами признаков, имеет вид:

$$P(t) = (1-t) * P_1 + t * (P_2 - d * N), \quad (1)$$

где:

d – семантическая дистанция; N – нормализованный вектор направления; t – параметр интерполяции; P_1, P_2 – начальная и конечная точки.

Используя значение семантической дистанции и метод интерполяции (способ нахождения промежуточных значений величины по имеющемуся дискретному набору известных значений) оптимизированный метод формализации можно представить в виде следующей формулы (формула 2):

$$\sum_{i=0}^{n} \frac{(N_a + N_b + ... + N_n)}{P(t)} * L_n(x), \quad (2)$$

где:

N_a,N_b,N_n – семантические единицы; $P(t)$ – семантическая дистанция; $L_n(x)$ – множитель Лагранжа.

Применение данного метода вызвано необходимостью повышения качества и репрезентативности наборов данных на этапе их предобработки, также необходимостью объективной метрики для оценки и корректировки признакового пространства.

Заключение

В основе метода лежат четыре этапа подготовки данных: очистка, нормализация, трансформация и балансировка. Поэтапная реализация этих операций повышает качество входных данных и снижает риск переобучения нейронных сетей.

Ключевой элемент метода - математическая формализация на основе семантической интерполяции. Интерполяция выполняется между точками в скрытом пространстве с учетом семантической дистанции и направленных векторных представлений. Данный механизм сохраняет семантическую целостность информации и оказывается особенно полезным при работе со сложными данными в неявных моделях.

Модуль встраивает описанный формальный метод в сквозной управляемый процесс. В рамках этого процесса решаются две задачи: оптимизация структуры данных (выбор типов полей, например VARCHAR для строковых значений) и внедрение механизмов контроля качества (логирование изменений, промежуточные отчеты). Благодаря этому обеспечиваются воспроизводимость и прозрачность обработки.

Разработанная модель данных реализована в СУБД MariaDB. Она создает основу для хранения метаинформации и поддерживает дальнейшие разработки в области автоматизированной обработки. Метод позволяет автоматизировать подготовку датасетов, сократить временные затраты и стандартизировать процесс для разных типов задач. Перспективное направление - интеграция рекомендательных систем для оптимизации гиперпараметров обработки; это может привести к созданию адаптивных решений.

Выделяются три основных направления для эволюции подхода.

Первое направление - адаптация архитектуры под легковесные модели. Это обусловлено ростом числа IoT-устройств и потребностью в локальной обработке данных с ограниченными ресурсами.

Второе направление - внедрение дополненной аналитики для автоматизации формирования гипотез, что упрощает работу специалистов без глубоких технических знаний.

Третье направление - разработка соответствующего программного модуля. Ключевые задачи этого модуля: интеграция с разнородными источниками данных (включая потоковые и слабоструктурированные) и реализация методов активного обучения.

В качестве расширения функционала предполагается включение алгоритмов автоматического подбора методов предобработки с анализом их влияния на результаты обучения. Это соответствует современным трендам - автоматическому машинному обучению (AutoML) и объяснимому искусственному интеллекту (XAI).

Для обработки текстовых данных перспективной признается интеграция контекстуализированных векторных представлений слов (например, BERT или ELMo). Такая интеграция повышает качество многоязычной обработки (Multilingual NLP) и расширяет возможности модуля для международных проектов.

Реализация перечисленных шагов подтверждает, что качественная и адаптивная обработка данных служит необходимым фундаментом для успешного применения нейронных сетей - как в промышленных, так и в научных проектах.

Литература:
1. Kotsiantis S., Kanellopoulos D., Pintelas P. Data preprocessing for supervised leaning // International Journal of Computer Science. 2006. Vol. 1. No. 1. P. 111–117.
2. Sculley D., Holt G., Golovin D., et al. Hidden technical debt in machine learning systems. In: Advances in Neural Information Processing Systems 28 (NeurIPS 2015). Proceedings.com, 2015. P. 2503–2511.
3. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). Minneapolis, MN, USA, 2019. Vol. 1. P. 4171–4186.
4. Sennrich R., Haddow B., Birch A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016. Vol. 1: Long Papers. P. 1715–1725.
5. García S., Ramírez-Gallego S., Luengo J., et al. Big data preprocessing: methods and prospects // Big Data Analytics. 2016. Vol. 1. No. 1. Art. no. 9.
6. Ravaglia L., Rusci M., Nadalini D., et al. A TinyML platform for on-device continual learning with quantized latent replays. arXiv. arXiv, 2021.
7. Zhang Y., Liu Q., Song L. Sentence-state LSTM for text representation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). Association for Computational Linguistics, 2018. P. 317–327.
8. Qiu J., Wang B., Zhou C., et al. Pre-trained models for natural language processing: A survey // Science China Technological Sciences. 2020. Vol. 63. No. 10. P. 1872–1897.
9. Буль И.В., Антонова К.Н. К вопросу о классификации мультимодальных текстов нового порядка. В сб.: Диалог поколений: Учиться. Учить. Изучать: труды V Всероссийской научно-практической конференции с международным участием: в 2 ч. Санкт-Петербург: Высшая школа технологии и энергетики СПбГУПТД, 2024. С. 59–62. [Bul I.V., Antonova K.N. On the issue of classification of multimodal texts of a new order. In: Dialogue of generations: Learn. Teach. Study: Proceedings of the V All-Russian Research and Practice Conference with International Participation: in 2 parts. Saint Petersburg: Higher School of Technology and Energy, 2024. P. 59–62 (in Russian)].
10. Бурлаева Е.И. Обзор методов классификации текстовых документов на основе подхода машинного обучения // Программная инженерия. 2017. Т. 8. № 7. С. 328–336. [Burlaeva E.I. Review of text document classification methods based on machine learning approach // Software Engineering. 2017. Vol. 8. No. 7. P. 328–336 (in Russian)].
11. Андреева А.Е. Методы обработки естественного языка (NLP): от классических моделей до больших языковых моделей (LLM) // Академическая публицистика. 2025. № 12-2. С. 79–83. [Andreeva A.E. Natural language processing (NLP) methods: from classical models to large language models (LLMs) // Academic Journalism. 2025. No. 12-2. P. 79–83 (in Russian)].

Благодарности:
Работа поддержана Министерством науки и высшего образования Российской Федерации в рамках базовой части государственного задания для высших заведений #FRRR 2026-0006.

Об авторах:
АНТОНОВ Вячеслав Викторович, профессор кафедры Автоматизированных систем управления, Уфимский университет науки и технологий, antonov.v@bashkortostan.ru.
АЛЕКСЕЕВА Дарья Сергеевна, аспирант кафедры Автоматизированных систем управления, Уфимский университет науки и технологий, ads.stat@mail.ru.
КОНОНОВ Никита Алексеевич, аспирант кафедры Автоматизированных систем управления, Уфимский университет науки и технологий, knnv.nkt@gmail.com.
ЮЛИНА Елизавета Вениаминовна, аналитик ООО «Центр информационных технологий», liza.yulina.02@mail.ru.

ДЬЯЧКОВ Алексей Сергеевич, руководитель направления Дирекции инновационного развития Департамента технологий и клиентских решений Управления технологического развития, ООО «Газпромнефть-Цифровые решения», dyachkov.ase@gazprom-neft.ru

Metadata:

Title: Development of an automated text data preprocessing module with adaptive control for neural network models.

Author 1: Vyacheslav Viktorovich Antonov, Professor of the Department of Automated Control Systems, Ufa State University of Science and Technology, 32 Zaki Vali st., 450076 Ufa, Republic of Bashkortostan, Russia, antonov.v@bashkortostan.ru.

Author 2: Darya Sergeevna Alekseeva, Postgraduate student of the Department of Automated Control Systems, Ufa University of Science and Technology, 32 Zaki Vali st., 450076 Ufa, Republic of Bashkortostan, Russia, ads.stat@mail.ru.

Author 3: Nikita Alexeyevich Kononov, Postgraduate student of the Department of Automated Control Systems, Ufa State University of Science and Technology, 32 Zaki Vali st., 450076 Ufa, Republic of Bashkortostan, Russia, knnv.nkt@gmail.com.

Author 4: Elizaveta Veniaminovna Yulina, analyst, LLC «Information Technology Center», 70/1 Revolutsionnaya st., 450005 Ufa, Republic of Bashkortostan, Russia, liza.yulina.02@mail.ru.

Author 5: Alexey Sergeevich Dyachkov, Head of the Department of Technological Development of Gazpromneft – Digital Solutions LLC, Bldg. 4, 5 Kievskaya st., 190013 Saint-Petersburg, Russia, dyachkov.ase@gazprom-neft.ru.

Abstract: The article discusses the development of an automated data processing system module for preparing datasets before training neural networks. The aim of the work is to overcome the problems associated with low quality and heterogeneity of input data, which lead to reduced productivity and retraining of models. The article proposes a comprehensive module architecture that includes a formal decomposition of the preprocessing process, a multi-level control loop for operational regulation and adaptation, as well as a detailed data model and database schema based on the MariaDB database management system. Special attention is paid to the methods of text information processing in NLP tasks. The presented solution makes it possible to standardize and significantly speed up the data preparation stage, minimizing manual labor and the influence of the human factor. The results of the work will be useful to researchers and developers seeking to improve the quality, reproducibility and efficiency of machine learning models through reliable automation of data preprocessing.

Keywords: Automated data processing, data preprocessing, neural networks, machine learning, natural language processing.