МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математической кибернетики и компьютерных наук

РАЗРАБОТКА И ВНЕДРЕНИЕ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ВЕБ-ПЛАТФОРМЫ «ЭЛЕКТРОННЫЙ КОРРЕКТОР»

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 411 группы
направления 02.03.02 — Фундаментальная информатика и информационные технологии
факультета компьютерных наук и информационных технологий
Просвириной Алины Александровны

Научный руководитель
профессор, д. т. н., доцент _____ А. С. Богомолов

Заведующий кафедрой
к. ф.-м. н., доцент _____ С. В. Миронов

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы. В настоящее время в связи с ростом количества информации в учебном и научном процессе существует потребность в повышении степени автоматизации нормоконтроля учебных и научных работ. Полная ручная проверка документов требует значительного времени сотрудников и может сопровождаться неучетом ряда ошибок в силу субъективных причин. Дополнительная сложность заключается в том, что требования к оформлению часто содержатся в ГОСТ, локальных методических указаниях и других нормативных PDF-документах, где они представлены на естественном языке, содержат пояснения, примеры и служебную информацию.

Существующие программные решения обычно решают отдельные задачи: помогают оформлять библиографические ссылки, проверяют грамотность текста, оценивают оригинальность работы или предоставляют шаблоны документов. Однако такие средства не позволяют автоматически извлекать требования из нормативных документов и применять их для комплексной проверки оформления. Поэтому разработка интеллектуального модуля, который способен выделять требования, классифицировать их и использовать при автоматизированной проверке, является актуальной задачей.

Для решения этой задачи в рамках программы «Стартап как диплом» разрабатывается веб-платформа «Электронный корректор». Она предназначена для автоматизированной проверки документов на соответствие требованиям ГОСТ и локальных методических материалов. В данной бакалаврской работе рассматривается ML-модуль платформы, отвечающий за обработку нормативных PDF-документов, формирование наборов правил и их последующее применение при проверке пользовательских работ.

Цель бакалаврской работы — разработать и внедрить ML-модуль платформы «Электронный корректор», обеспечивающий извлечение, классификацию и структурирование требований к оформлению документов, а также применение полученных наборов правил при автоматизированной проверке пользовательских работ.

Поставленная цель определила следующие задачи:

- проанализировать задачу автоматизированного нормоконтроля документов, существующие решения и роль ML-модуля в составе платформы;
- исследовать нормативные PDF-документы как источник требований и

определить формат входных и выходных данных модуля;

- подготовить обучающие данные и обучить модели машинного обучения для выявления требований и их классификации;
- реализовать обработку PDF-документов, извлечение текстовых фрагментов, формирование и нормализацию правил;
- интегрировать ML-модуль с backend-частью платформы и реализовать применение сохраненных наборов правил при проверке пользовательских документов;
- провести тестирование разработанного модуля на тестовых данных и практических сценариях.

Методологические основы работы составляют методы машинного обучения и обработки естественного языка, подходы к классификации текстовых фрагментов, технологии извлечения данных из PDF-документов, а также средства серверной разработки и интеграции программных компонентов.

Теоретическая значимость бакалаврской работы заключается в анализе задачи автоматизированного извлечения требований из нормативных документов, рассмотрении методов машинного обучения для классификации текстовых данных и формализации структуры правил, используемых при проверке оформления.

Практическая значимость бакалаврской работы заключается в разработке ML-модуля, который может использоваться в составе веб-платформы «Электронный корректор» для обработки нормативных PDF-документов, формирования наборов правил и автоматизированной проверки пользовательских работ.

Структура и объем работы. Бакалаврская работа состоит из введения, 2 разделов, заключения, списка использованных источников и 3 приложений. Общий объем работы — 52 страницы, включая 6 рисунков, 2 таблицы, список использованных источников из 21 наименования и приложения с исходным кодом программы.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Проектирование ML-модуля веб-платформы автоматизированной проверки документов» включает анализ предметной области, обзор существующих решений, рассмотрение методов машинного обучения в задачах анализа текста и формализацию данных, с которыми работает ML-модуль.

Анализ предметной области. В работе рассматривается задача автоматизированного нормоконтроля учебных и научных документов. Требования к оформлению задаются ГОСТ, локальными документами университета и методическими указаниями кафедр. Такие документы часто представлены в формате PDF, а сами требования записаны на естественном языке. Из-за этого перед автоматической проверкой необходимо выделить из текста только те фрагменты, которые действительно являются правилами оформления, и привести их к виду, удобному для дальнейшего использования системой.

Существующие решения для подготовки и проверки документов. В работе рассмотрены инструменты Zotero, Google Scholar, LanguageTool, «Антиплагиат», Overleaf и LaTeX-шаблоны. Анализ показал, что эти решения полезны на отдельных этапах подготовки учебной или научной работы: при поиске литературы, оформлении ссылок, проверке текста или работе с шаблонами. При этом они не решают задачу автоматического извлечения требований из нормативных документов и не позволяют формировать на их основе наборы правил для комплексной проверки оформления.

Методы машинного обучения в задачах анализа текста. Для решения поставленной задачи рассматриваются методы обработки естественного языка и классификации текстовых фрагментов. Особое внимание уделяется моделям семейства BERT и RuBERT, которые позволяют учитывать контекст слов в русскоязычном тексте. В рамках ML-модуля решаются две связанные задачи: определение того, является ли фрагмент текста требованием к оформлению, и классификация найденного требования по тематической категории.

Формализация входных и выходных данных ML-модуля. Входными данными ML-модуля является нормативный PDF-документ, содержащий требования к оформлению. В результате обработки формируется структурированный набор правил. Каждое правило содержит текст требования, категорию, связь с исходным документом и служебные данные, необходимые для хранения, отображения и дальнейшего применения при проверке пользовательских работ.

Второй раздел «Разработка и тестирование ML-модуля платформы “Электронный корректор”» посвящен практической реализации модуля: подготовке обучающих данных, обучению моделей, обработке PDF-документов, нормализации правил, интеграции с backend-частью платформы и тестированию.

Разработка моделей машинного обучения и подготовка обучающих данных. Для обучения моделей были подготовлены датасеты, содержащие текстовые фрагменты нормативных документов и соответствующие метки. Первая модель предназначена для определения требований в тексте, вторая — для классификации найденных требований по категориям. Категории связаны с основными элементами оформления документа: структурой работы, таблицами, рисунками, списком использованных источников, параметрами шрифта и другими требованиями.

Обучение и оценка качества моделей. В качестве основы использовалась модель DeepPavlov/rubert-base-cased, адаптированная для задач классификации текстовых фрагментов. Обучающие данные были разделены на обучающую, валидационную и тестовую выборки. Для оценки качества применялись метрики accuracy, precision, recall и F1-score. По результатам тестирования модель определения правил показала значение F1-score, равное 0.9901, а модель классификации категорий — значение macro F1-score, равное 0.9918. Полученные результаты подтверждают высокое качество работы моделей на тестовой выборке.

Обработка нормативных PDF-документов. Разработанный модуль выполняет извлечение текста из загруженного PDF-файла, разбиение текста на фрагменты и передачу этих фрагментов в модель определения правил. После этого найденные требования классифицируются по категориям. Для отбора результатов используется порог уверенности модели, что позволяет исключать нерелевантные фрагменты и повышать качество итогового набора правил.

Формирование и хранение набора правил. После обработки PDF-документа найденные требования проходят нормализацию. На этом этапе устраняются дубли, сохраняется связь с исходным документом, уточняется категория правила и формируется единый формат хранения. Сформированный набор правил может быть повторно использован при проверке пользовательских документов, что позволяет платформе работать не только с заранее заданными требованиями, но и с требованиями, извлеченными из загружаемых методиче-

ских материалов.

Интеграция ML-модуля с платформой. ML-модуль был интегрирован с backend-частью веб-платформы «Электронный корректор». Для взаимодействия с ним реализован API-маршрут, который принимает PDF-файл, запускает обработку документа, проверяет доступность обученных моделей, формирует набор правил и сохраняет результат. Благодаря этой интеграции извлеченные требования могут применяться при автоматизированной проверке пользовательских работ.

Тестирование ML-модуля. Работоспособность модуля была проверена на тестовых данных и практических сценариях. В ходе тестирования были продемонстрированы извлечение правил из нормативного документа через Swagger, обнаружение нарушений в списке использованных источников, выявление отсутствующих структурных элементов работы, определение нарушений требований к оформлению шрифта, а также проверка одной и той же работы по разным наборам правил. Результаты подтвердили корректность работы реализованного подхода.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы был разработан и внедрен ML-модуль веб-платформы «Электронный корректор», предназначенный для извлечения, классификации и структурирования требований к оформлению документов, а также применения полученных правил при автоматизированной проверке пользовательских работ.

Перед началом разработки была проанализирована предметная область автоматизированного нормоконтроля документов и рассмотрены существующие решения, используемые при подготовке и проверке учебных и научных работ. Анализ показал, что существующие инструменты решают отдельные задачи, но не обеспечивают автоматическое извлечение требований из нормативных PDF-документов и их использование при комплексной проверке оформления.

Для реализации проекта были использованы методы машинного обучения и обработки естественного языка. Были подготовлены обучающие данные и обучены две модели: модель определения требований в тексте нормативного документа и модель классификации найденных требований по категориям. Использование RuBERT позволило учитывать особенности русскоязычных текстов и достичь высоких значений метрик качества.

В процессе разработки был реализован программный модуль, выполняющий обработку PDF-документов, извлечение текстовых фрагментов, определение требований, классификацию, нормализацию и сохранение сформированных правил. Также была выполнена интеграция ML-модуля с backend-частью платформы, что позволило применять сохраненные наборы правил при проверке пользовательских документов.

Результаты тестирования показали, что система корректно извлекает, классифицирует и сохраняет требования, а также применяет выбранные наборы правил для выявления нарушений в пользовательских документах. Проверка практических сценариев подтвердила работоспособность разработанного модуля и возможность его использования в составе веб-платформы «Электронный корректор».

Разработанный ML-модуль имеет практическую значимость, поскольку позволяет сократить время нормоконтроля, снизить влияние человеческого фактора и расширить возможности автоматизированной проверки документов. Проект «Электронный корректор» выполнялся в рамках программы «Стартап

как диплом» и получил положительную экспертную оценку.

Таким образом, задачи выпускной квалификационной работы решены и цель ее достигнута. Разработанный модуль может быть использован как основа для дальнейшего развития платформы, расширения набора поддерживаемых требований и повышения качества автоматизированной проверки учебных и научных работ.