МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ПРОВЕРКИ СОЧИНЕНИЙ ЕГЭ ПО РУССКОМУ ЯЗЫКУ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 273 группы
направления 02.04.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Суховой Екатерины Владимировны

Научный руководитель:
старший преподаватель _____ Лапшева Е.Е.

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ Огнева М. В.

подпись, дата

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы. В современном мире стремительно развиваются технологии искусственного интеллекта, в частности, обработки естественного языка (NLP, англ. Natural Language Processing). Эти технологии применяются в различных сферах деятельности человека, включая образование [1]. Одним из важных направлений цифровизации образовательного процесса является разработка автоматизированных систем для проверки знаний учащихся. Существует большое количество решений, которые позволяют облегчить и ускорить проверку знаний, однако до сих пор сложной задачей является оценка заданий, ответы на которые представлены в текстовой форме, таких как эссе и сочинения. Ручная проверка требует больших усилий и временных затрат со стороны учителя, помимо этого человек может пропустить ошибки в работе, а также на оценку может влиять субъективное мнение педагога.

В последние годы популярность набирают большие языковые модели, которые позволяют решать широкий круг задач в области взаимодействия с текстовыми данными. Так, модели хорошо справляются с генерацией и суммаризацией текстов, машинным переводом и другими задачами. Языковые модели обучены на огромных массивах текстов, что позволяет использовать их для работы с различными форматами текстов на различных языках. В частности, их используют для оценки эссе и сочинений [2].

Задача автоматизированной проверки текстовых работ приобретает особое значение в контексте Единого государственного экзамена (ЕГЭ) по русскому языку, обязательной частью которого является написание сочинения по предложенному тексту. Большое количество критериев оценивания и значительный объём проверяемых работ делают ручную проверку сочинений ЕГЭ одной из наиболее ресурсоёмких задач. Кроме того, специфика заданий с открытым ответом создаёт необходимость согласования оценок между несколькими проверяющими, что приводит к увеличению временных затрат на проверку.

Цель магистерской работы – разработать автоматизированную систему оценки сочинений на русском языке в формате ЕГЭ согласно предоставленным критериям.

Поставленная цель определила следующие **задачи**:

1. Изучить предметную область и существующие аналоги в области анализа и оценки текста.
2. Рассмотреть методы и средства сбора данных, необходимые для формирования датасета, и найти подходящие источники с текстами на русском языке.
3. Сформировать датасет для обучения моделей машинного обучения.
4. Найти большие языковые модели, подходящие для анализа текста на русском языке.
5. Исследовать стратегии борьбы с дисбалансом классов.
6. Дообучить большие языковые модели на собранных данных и провести сравнительный анализ результатов обучения.
7. Разработать пользовательский интерфейс для загрузки текстов и вывода развёрнутых результатов оценки.
8. Провести апробацию разработанной системы на выборке сочинений с экспертной оценкой.

Методологические основы. Теоретическую базу исследования составили работы в области архитектуры трансформера и моделей на основе кодировщиков – А. Васвани [13], Дж. Девлина [6], И. Лю [14]; в области многозадачного обучения – Р. Каруаны [24]; в области стратегий борьбы с дисбалансом классов – Т. Линя [28] и Н. В. Чавла [29]; а также исследования по автоматизированной оценке письменных работ с помощью больших языковых моделей – А. Пака [15] и М. Лундгрена [2].

Практическая значимость магистерской работы заключается в создании работающего инструмента автоматизированной оценки сочинений ЕГЭ по русскому языку по двенадцати критериям. В отличие от

существующих русскоязычных сервисов проверки текстов, разработанная система оценивает не только грамотность, но и содержательные аспекты сочинения. Предложенная методика генерации синтетических обучающих данных из эталонных сочинений может быть применена и в других задачах классификации текстов в условиях критического дисбаланса классов.

Структура и объём работы. Магистерская работа состоит из введения, трёх разделов, заключения, списка использованных источников и шести приложений. Общий объём работы – 128 страниц, из них основное содержание – 79 страниц, включая 5 рисунков и 15 таблиц; список использованных источников информации – 52 наименования.

КРАТКОЕСОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Большие языковые модели и автоматизированная оценка текстов» посвящён обзору больших языковых моделей и их применения для автоматизированной оценки письменных работ.

Рассматриваются три основных типа архитектур современных языковых моделей: модели на основе архитектуры трансформера, модели, состоящие из композиции кодировщиков, и генеративные модели на основе декодировщика. Подробно разбирается архитектура трансформера с механизмом внимания, а также модели на основе кодировщиков – BERT и RoBERTa, – которые благодаря двунаправленному контексту и парадигме «pre-train – fine-tune» наиболее подходят для задач классификации и семантического анализа текстов, особенно при ограниченном объёме размеченных данных.

Отдельно рассматривается задача автоматизированной проверки письменных работ. Приводятся исследования, показывающие, что большие языковые модели существенно повысили качество автоматической оценки эссе, однако для получения надёжных результатов требуется их гибкая настройка и дообучение на данных, близких к оцениваемым работам. В обзоре существующих аналогов проанализированы русскоязычные сервисы «Орфограммка», «Тургенев» и «Главред», которые ориентированы преимущественно на проверку грамотности и стиля. Сделан вывод, что задача автоматизированной оценки сочинений ЕГЭ по русскому языку не имеет готового решения среди существующих систем.

Второй раздел «Методы машинного обучения для классификации текстов» посвящён формализации задачи и методам её решения.

Рассмотрены формат и структура сочинения ЕГЭ, а также проведена формализация критериев оценивания в задачи машинного обучения. Каждый из двенадцати критериев К1–К12 соотнесён с типом задачи – бинарной классификации, многоклассовой классификации или регрессии. Показано, что автоматизированная оценка сочинения представляет собой комплексную

многозадачную проблему, в которой ряд критериев требует семантического анализа пары «исходный текст – сочинение», а ряд критериев – анализа структуры и грамотности.

Описан подход многозадачного обучения с общим кодировщиком и набором независимых классификационных голов, его преимущества (регуляризация, единый инференс) и ограничения, главным из которых является негативный перенос между разнородными задачами. Рассмотрены основные стратегии борьбы с дисбалансом классов: взвешивание функции потерь, Focal Loss, oversampling класса меньшинства, настройка порога классификации, генерация синтетических данных и ансамблевый подход (голосование и стекинг). Отдельно описаны метрики оценки качества классификации; в качестве основной метрики в условиях дисбаланса обоснован выбор macro F1, а в качестве вспомогательных – minority F1 и матрица ошибок.

Третий раздел «Экспериментальное исследование и реализация» содержит подробное описание практической части работы.

Все этапы выполнялись на локальной машине с графическим процессором с поддержкой технологии CUDA. Сбор данных проводился методом веб-скрапинга с трёх образовательных онлайн-ресурсов – «Могу Писать», «4ЕГЭ» и «Литрекон» – с использованием библиотек requests, BeautifulSoup и pandas. В результате сбора и предобработки сформирован основной датасет из 2117 сочинений с экспертной разметкой по двенадцати критериям, а также отдельный корпус эталонных сочинений. Разведочный анализ выявил ограниченный объём данных, выраженный дисбаланс классов для ряда критериев (К1, К7, К11) и ограничение длины входной последовательности, что и определило выбор стратегий обучения.

Для экспериментов отобраны три модели на основе кодировщиков с поддержкой русского языка: rubert-tiny2 (29 млн параметров), rubert-base-cased (178 млн) и ruRoberta-large (355 млн). Исследованы следующие подходы: многозадачное обучение в единой конфигурации;

специализированные модели по группам критериев («семантика», «грамотность», «структура»); комбинация Focal Loss и oversampling; настройка порога классификации; генерация синтетических данных и ансамблевый подход.

Установлено, что при ограниченном объёме данных компактные модели предпочтительнее крупных для содержательных критериев, тогда как для критериев грамотности крупная модель ruRoberta-large раскрывает свой потенциал только в сочетании с Focal Loss и oversampling. Специализированные модели по группам критериев превзошли единую модель для 9 критериев из 12, что подтверждает наличие негативного переноса. Наибольший прирост качества обеспечила генерация синтетических данных из эталонных сочинений путём контролируемого внесения ошибок: для критерия К7 minority F1 вырос с 0,093 до 0,628, для критерия К11 – с 0,167 до 0,759. Ансамблевый метод голосования дал точечные улучшения для отдельных критериев, а стекинг при имеющемся объёме данных оказался неэффективным. Средний macro F1 по лучшим конфигурациям для каждого критерия составил 0,557.

Для практического применения системы реализован пользовательский интерфейс в виде веб-приложения на платформе Streamlit, объединяющий лучшие обученные модели в единый инструмент. Приложение поддерживает два режима работы, обеспечивает ввод сочинения текстом или загрузку из файла и формирует наглядное представление результатов в виде графика с цветовой кодировкой по критериям и таблиц с детализацией оценок.

Для оценки практической применимости проведена апробация системы на независимой выборке из 20 сочинений с опубликованной экспертной оценкой. Среднее абсолютное отклонение итогового балла от экспертной оценки составило 1,95 балла; у 70 % сочинений отклонение не превысило двух баллов, у 90 % – трёх баллов. Это позволяет рассматривать систему как вспомогательный инструмент предварительной оценки, но не как замену экспертной проверки.

ЗАКЛЮЧЕНИЕ

В данной работе была разработана автоматизированная система оценки сочинений на русском языке в формате ЕГЭ по двенадцати критериям К1–К12 на основе дообученных языковых моделей-кодировщиков. Были рассмотрены основные архитектуры современных языковых моделей, методы многозадачного обучения и стратегии борьбы с дисбалансом классов.

С помощью веб-скрапинга трёх образовательных онлайн-ресурсов был сформирован набор из 2117 сочинений с экспертной разметкой по всем двенадцати критериям, а также отдельный корпус эталонных сочинений, использованных в дальнейшем для генерации синтетических данных.

Было исследовано несколько подходов к обучению моделей, с учетом основной проблемы – дисбаланса классов. В частности: многозадачное обучение в единой конфигурации, специализированные модели по группам семантически близких критериев, применение Focal Loss совместно с oversampling, настройка порога классификации, генерация синтетических данных и применение ансамблей. По итогам обучения трех моделей – *rubert-tiny2*, *rubert-base-cased* и *ruRoberta-large* – средний macro F1 по лучшим конфигурациям составил 0,557. Было установлено, что для содержательных критериев предпочтительны компактные модели, а для критериев грамотности – крупные модели при условии применения Focal Loss и oversampling. Специализированные модели по группам критериев превзошли единую модель для 9 критериев из 12.

Главным методическим вкладом работы является методика генерации синтетических обучающих данных из эталонных сочинений посредством контролируемого внесения ошибок заданного типа. Подход оказался решающим для критериев с критическим дисбалансом: для критерия К7 F1 класса меньшинства вырос с 0,093 до 0,628, для критерия К11 – с 0,167 до 0,759. Этот метод оказался наиболее эффективным из всех исследованных способов борьбы с дисбалансом классов и существенно превзошёл стандартные стратегии – взвешивание функции потерь, oversampling и Focal

Loss.

Для практического применения системы был реализован веб-интерфейс на платформе Streamlit, объединяющий лучшие обученные модели в единый инструмент. Интерфейс поддерживает два режима работы, обеспечивает ввод сочинения текстом или загрузку из файла, формирует наглядное представление результатов в виде графика с цветовой кодировкой по критериям и таблиц с детализацией оценок. В отличие от существующих русскоязычных сервисов проверки текстов, разработанная система оценивает не только грамотность, но и содержательные аспекты сочинения. Для оценки практической применимости системы была проведена апробация на независимой выборке из 20 сочинений: среднее абсолютное отклонение итогового балла от экспертной оценки составило 1,95 балла, у 70 % сочинений отклонение не превысило двух баллов.

В результате получаем систему, которая на данный момент не сможет заменить эксперта, но при этом может являться вспомогательным инструментом для примерной оценки сочинения.

Промежуточные и основные результаты работы были представлены на следующих научных мероприятиях:

1. Всероссийская научно-практическая конференция «Информационные технологии в образовании» (7–8 ноября 2025 года); по итогу участия статья «Разработка датасета для автоматической проверки сочинений ЕГЭ по русскому языку: методы сбора и обработки данных» опубликована в сборнике трудов конференции.
2. Студенческая научная конференция факультета КНиИТ СГУ (23 апреля 2026 года); доклад на тему «Разработка автоматизированной системы проверки сочинений на русском языке в формате ЕГЭ».

Основные источники информации

1. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HTL. – 2019. – Vol. 1. – pp. 4171–4186. – DOI: 10.18653/v1/N19-1423.

2. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – pp. 5998–6008.

3. Liu Y., Ott M., Goyal N. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach // arXiv preprint. – 2019. – No. 1907.11692. – DOI: 10.48550/arXiv.1907.11692.

4. Caruana R. Multitask Learning // Machine Learning. – 1997. – No. 28. – pp. 41–75.

5. Lin T.-Y., Goyal P., Girshick R., He K., Dollár P. Focal Loss for Dense Object Detection // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2018. – Vol. 42, No. 2. – pp. 318–327. – DOI: 10.1109/TPAMI.2018.2858826.

6. Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. SMOTE: Synthetic Minority Over-sampling Technique // Journal of Artificial Intelligence Research. – 2002. – No. 16. – pp. 321–357.

7. Pack A., Barrett A., Escalante J. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability // Computers and Education: Artificial Intelligence. – 2024. – Vol. 6, No. 100234. – DOI: 10.1016/j.caeai.2024.100234.

8. Lundgren M. Large Language Models in Student Assessment: Comparing ChatGPT and Human Graders // arXiv preprint. – 2024. – No. 2406.16510. – DOI: 10.13140/RG.2.2.27630.42561.