MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 32, NO. 3, 2025
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Modern Russian-language Texts Models Comparison for the Task of CEFR Levels Classification

V. A. Lavrovskiy[1], N. S. Lagutina[1], O. B. Lavrovskaya[1]

DOI: 10.18255/1818-1015-2025-3-298-310

[1]P.G. Demidov Yaroslavl State University, Yaroslavl, Russia

MSC2020: 68T50
Research article
Full text in Russian

Received August 4, 2025
Revised August 25, 2025
Accepted August 27, 2025

The development of high-quality tools for automatic determination of text levels according to the CEFR scale allows creating educational and testing materials more quickly and objectively. In this paper, the authors examine two types of modern text models: linguistic characteristics and embeddings of large language models for the task of classifying Russian-language texts by six CEFR levels: A1–C2 and three broader categories A, B, C. The two types of models explicitly represent the text as a vector of numerical characteristics. In this case, dividing the text into levels is considered as a common classification task in the field of computational linguistics. The experiments were conducted with our own corpus of 1904 texts. The best quality is achieved by rubert-base-cased-conversational without additional adaptation when determining both six and three text categories. The maximum F-measure value for levels A, B, C is 0.77. The maximum F-measure value for predicting six text categories is 0.67. The quality of text level determination depends more on the model than on the machine learning classification algorithm. The results differ from each other by no more than 0.01–0.02, especially for ensemble methods.

Keywords: natural language processing; Russian-language texts classification; linguistic characteristics; embeddings; BERT; GPT; CEFR

INFORMATION ABOUT THE AUTHORS

Lavrovskiy, Vadim A.	ORCID iD: 0000-0003-3147-077X. E-mail: comado19@gmail.com Post-graduate Student
Lagutina, Nadezhda S. (corresponding author)	ORCID iD: 0000-0002-6137-8643. E-mail: lagutinans@rambler.ru PhD, Associate Professor
Lavrovskaya, Olga B.	ORCID iD: 0009-0000-1575-7098. E-mail: o.lavrovskaya@uniyar.ac.ru PhD, Associate Professor

Funding: The work was carried out within the framework of a grant from the Ministry of Social Communications and Scientific and Technological Development of the Yaroslavl Region on the topic of scientific research "Automatic text analysis", agreement No. 17NP/2024 dated December 24, 2024.

For citation: V. A. Lavrovskiy, N. S. Lagutina, and O. B. Lavrovskaya, "Modern Russian-language texts models comparison for the task of CEFR levels classification", *Modeling and Analysis of Information Systems*, vol. 32, no. 3, pp. 298–310, 2025.
DOI: 10.18255/1818-1015-2025-3-298-310.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 32, № 3, 2025
сайт журнала: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Сравнение современных моделей русскоязычных текстов для задачи классификации по уровням CEFR

В. А. Лавровский[1], Н. С. Лагутина[1], О. Б. Лавровская[1]

DOI: 10.18255/1818-1015-2025-3-298-310

[1]Ярославский государственный университет им. П.Г. Демидова, Ярославль, Россия

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 4 августа 2025 г.
После доработки 25 августа 2025 г.
Принята к публикации 27 августа 2025 г.

Разработка качественных инструментов автоматического определения уровней текстов по шкале CEFR позволяет создавать учебные и проверочные материалы более быстро и объективно. В данной работе авторы исследуют два типа современных моделей текста: лингвистические характеристики и эмбеддинги больших языковых моделей для задачи классификации русскоязычных текстов по шести уровням CEFR: A1–C2 и трём укрупнённым категориям A, B, C. Два вида моделей явным образом представляет текст в виде вектора числовых характеристик. При этом разделение текста на уровни рассматривается как обычная задача классификации в области компьютерной лингвистики. Эксперименты проводились с собственным корпусом из 1904 текстов. Лучшее качество достигается rubert-base-cased-conversational без дополнительной адаптации при определении как шести, так и трёх категорий текста. Максимальное значение F-меры для уровней A, B, C равно 0,77. Максимальное значение F-меры для прогнозирования шести категорий текста равно 0,67. Качество определения уровня текста больше зависит от модели, чем от алгоритма классификации машинного обучения. Результаты отличаются друг от друга не более чем на 0,01–0,02, особенно это касается ансамблевых методов.

Ключевые слова: автоматическая обработка текста; классификация русскоязычных текстов; лингвистические характеристики; эмбеддинги; BERT; GPT; CEFR

ИНФОРМАЦИЯ ОБ АВТОРАХ

Лавровский, Вадим Алексеевич	ORCID iD: 0000-0003-3147-077X. E-mail: comado19@gmail.com Аспирант
Лагутина, Надежда Станиславовна (автор для корреспонденции)	ORCID iD: 0000-0002-6137-8643. E-mail: lagutinans@rambler.ru Канд. физ.-мат. наук, доцент
Лавровская, Ольга Борисовна	ORCID iD: 0009-0000-1575-7098. E-mail: o.lavrovskaya@uniyar.ac.ru Канд. пед. наук, доцент

Финансирование: Работа выполнена в рамках гранта министерства социальных коммуникаций и научно-технологического развития Ярославской области по теме научного исследования «Автоматический анализ текстов», соглашение № 17НП/2024 от 24 декабря 2024 года.

Для цитирования: V. A. Lavrovskiy, N. S. Lagutina, and O. B. Lavrovskaya, “Modern Russian-language texts models comparison for the task of CEFR levels classification”, *Modeling and Analysis of Information Systems*, vol. 32, no. 3, pp. 298–310, 2025. DOI: 10.18255/1818-1015-2025-3-298-310.

Введение

В области цифровизации обучения студентов и школьников важной проблемой является моделирование образовательных технологий и их последующая автоматизация. Решение этой задачи позволяет персонифицировать обучение, привить навыки самостоятельной работы, сформировать все виды компетенций образовательной программы, распространить лучшие педагогические методики на большое число учащихся, в том числе в области изучения естественных языков [1]. В прикладной лингвистике существует система общеевропейских компетенций владения иностранным языком (Common European Framework of Reference, CEFR). Эта система обеспечивает чёткую основу для разработки языковых учебных ресурсов и оценки уровня владения материалом [2]. В системе CEFR знания и умения учащихся делятся на три крупных категории: A — базовое владение языком, B — самодостаточное владение, C — свободное владение. Далее каждая категория делится на две части, всего получается шесть уровней: A1 — уровень выживания, A2 — предпороговый уровень, B1 — пороговый уровень, B2 — пороговый продвинутый уровень, C1 — уровень профессионального владения, C2 — уровень владения в совершенстве [3]. Разработка качественных инструментов автоматического определения уровней текстов является основой для создания учебных и проверочных материалов, так как позволяет сделать этот процесс более быстрым и объективным [4].

Интеграция и адаптация шкалы CEFR в российском образовании повышает качество и эффективность преподавания, так как дескрипторы этой шкалы ориентированы на оценку практических коммуникативных навыков владения языком. Единый стандарт и чётко проработанные критерии показывают обучающимся их достижения и ставят новые цели, а практическая направленность помогает смещать акценты обучения для разных специальностей, например, инженерных или гуманитарных [5]. Кроме того, автоматизация определения уровня текстов особенно актуальна для русского языка, где понимание доступности содержания существенно увеличивает качество учебных материалов [6]. Поэтому авторы работы поставили перед собой задачу исследовать два типа современных цифровых моделей для классификации русскоязычных текстов по уровням CEFR: лингвистические характеристики и эмбеддинги больших языковых моделей.

1. Обзор аналогичных научных исследований

В последние годы самыми популярными моделями в области автоматического анализа текста являются большие языковые модели на основе нейронных сетей. Числовые характеристики текста (эмбеддинги), полученные с их помощью, успешно моделируют текст в самых разных задачах. Классический метод предсказания одного из шести уровней CEFR англоязычных ответов учащихся с использованием эмбеддингов BERT представлен в работе [7]. Авторы дополнили исходный текст учащегося исправлениями, предоставляемыми автоматическим инструментом или людьми-оценщиками и рассмотрели способы, в которых тексты и исправления объединяются на входе модели BERT или путём слияния полученных заранее эмбедиингов BERT. Предлагаемый подход оценили на двух открытых наборах данных: базе данных открытого языка Кембриджского университета (EFCAMDAT) и корпусе Кембриджского университета для получения сертификата по английскому языку (CLC-FCE). Лучший результат для корпуса EFCAMDAT достиг очень высокого значения доли правильных ответов (accuracy) классификации: почти 0.98, однако для второго корпуса CLC-FCE оказался равен 0.62. Снижение качества классификации исследователи объясняют меньшим объёмом обучающего материала и менее качественной разметкой.

Анализ особенностей классификации англоязычных текстов на уровни и подуровни CEFR описан в статье [8]. Были использованы два открытых корпуса текстов: CEFR Levelled English Texts и BEA-2019. Классификация проводилась на все шесть категорий, на три уровня A, B и C и между отдельными подуровнями. Оказалось, что модели BERT показали качество предсказания в виде

F-меры равной 0.90 для трёх уровней, но всего 0.69 для шести. Исследование ошибок также подтвердило, что сложнее всего разделить между собой подуровни внутри категорий A, B и C.

Авторы работы [9] адаптировали модели BERT для классификации эссе по шести уровням CEFR на английском, французском и шведском языках. Они исследовали, как заморозка различных слоёв BERT во время обучения влияет на качество классификации, учитывая, что нижние слои выявляют базовые лингвистические признаки, а верхние слои — специфические характеристики, такие как семантические и контекстные признаки [10]. Интересно, что F-мера для классификации английских текстов равна 0.98, а для французских и шведских намного меньше: 0.57 и 0.74 соответственно. В другой статье [11], посвящённой национальному языку модель ArabicBERT показала F-меру равную 0.80 при определении уровня A, B или C для арабских текстов.

Детальное исследование определения уровня сложности разговорных текстов на английском языке провели Kogan и др. [12]. Авторы собрали собственный корпус из 990 текстов (Ace-CEFR) и поставили задачу прогнозировать метку текста по шкале от 1 до 6 в соответствии с уровнями CEFR. Основной метрикой качества была выбрана среднеквадратическая ошибка (MSE) между предсказанием модели и консенсусной экспертной оценкой. Были исследованы три типа моделей: линейная регрессия, BERT и большая языковая модель PaLM 2-L, для сравнения тестовый набор также был оценен экспертом-человеком. Среднеквадратическая ошибка разметок экспертов составила 0.75. MSE линейной регрессии с использованием характеристик, таких как длина предложений и слов, лексическое разнообразие, составила 0.81. MSE модели BERT оказалась равна 0.37, а PaLM 2-L — 0.48. Исследователи проанализировали ошибки и отметили, что модели, обученные на основе корпуса Ace-CEFR, измеряют сложность текста точнее, чем эксперты-люди, и работают достаточно быстро для промышленного использования.

Лексические и более сложные лингвистические характеристики часто используются для повышения качества решений. В работе [3] представлен метод классификации письменных работ (эссе) учащихся на английском языке в соответствии с шестью уровнями владения языком CEFR. Авторы оценивали значимость лингвистических характеристик и критериальных признаков для оценки уровня владения языком. Лучшие результаты показало сочетание всех типов характеристик, наибольшее значение доли правильных ответов достигло 0.82, среднее для нескольких корпусов равно 0.59. Исследователи показали универсальность модели для текстов из разных предметных областей и обратили внимание на практическую применимость решения в программных системах.

С помощью визуализации процедуры классификации дерева решений были выявлены лингвистические особенности, которые отличают образцы письменных работ учащихся разных уровней CEFR [13]. Исследователи обращают внимание, что такие характеристики могут быть использованы для автоматизации обратной связи с учащимися, отслеживания долгосрочного развития навыков письма, содействия экспертным оценкам письменных тестов.

Лингвистические характеристики играют важную роль при моделировании национальных языков. В статье [14] представлена модель автоматической классификации уровня владения письменной речью на итальянском языке как втором языке на четыре категории CEFR от B1 до C2. Авторы составили корпус из 3041 документа. Текст моделировался объединением лексических, морфосинтаксических и фразеологических характеристик. Лучший результат был получен с использованием классификатора случайного леса, F-мера достигла значения 0.89. Возможно высокое качество явилось следствием небольшого числа классов и особенностей собранного корпуса. К сожалению, для национальных языков большие размеченные корпуса часто отсутствуют, что приводит к снижению объективности результатов.

Модель определения уровня сложности русскоязычного текста по шкале CEFR из более чем 100 лингвистических признаков предложена в работе [15]. На её основе разработано программное

приложение «Текстометр». Модель обучена на коллекции из 800 текстов из пособий по русскому языку как иностранному. К сожалению, качество классификации не указано.

Glišić и др. [16] использовали различные лингвистические признаки, включая эмбеддинги, для разработки моделей автоматической классификации уровня владения языком по шкале CEFR для исландского языка. Авторы подчеркнули морфологическую сложность национального языка и ограниченность корпусов текстов. В работе сравнивались два подхода к оценке уровня текстов учащихся. Первый включал простые статистические характеристики текста на основе длины слов и лексического разнообразия, второй — как морфологические и лексические признаки, так и эмбеддинги IceBERT. В обоих вариантах использовались исходный и исправленный текст и учитывались ошибки/отклонения. Для первого подхода средняя F-мера оказалась в диапазоне 0.62–0.67, а для второго — 0.75–0.80.

Таким образом, определение уровня текста по шкале CEFR является интересной и важной задачей языкознания. В настоящее время разработано большое количество методов её автоматического решения преимущественно для английского языка. Однако, для развития образовательных ресурсов и распространения русского языка число исследований и степень их систематизации явно недостаточны.

2. Методика исследования

Анализ литературы позволяет выделить два основных способа представления текста: лингвистические характеристики и эмбеддинги больших языковых моделей. Лингвистические характеристики появились первыми в области автоматического анализа текстов, но не утратили своей актуальности, в первую очередь, для национальных языков и задач с ограниченными данными для обучения. Большие языковые модели показывают высокое качество классификации текстов по заданным категориям, так как успешно моделируют естественный язык, но чувствительны к размерам корпусов текстов, используемых в экспериментах, и качеству их разметки. Данное исследование проведено авторами с целью сравнения этих подходов к моделированию текстов в решении поставленной задачи.

2.1. Основные этапы метода

Авторами было проведено две группы экспериментов.

- Классификация с лингвистическими признаками:
  - расчёт лингвистических характеристик уровня слов, символов и структуры, формирование числовых векторов текстов;
  - обучение алгоритмов машинного обучения на полученных векторах;
  - классификация текстов по шести уровням (A1–C2) и трём (A, B, C).
- Классификация с эмбеддингами больших языковых моделей:
  - преобразование текстов в эмбеддинги с помощью предобученных языковых моделей;
  - обучение алгоритмов машинного обучения на эмбеддингах;
  - классификация текстов по шести уровням (A1–C2) и трём (A, B, C).

Для оценки качества определения уровня текста были использованы статистические показатели:

- точность — доля релевантных объектов среди всех объектов, которые алгоритм отнёс к заданному классу;
- полнота — доля релевантных объектов, которые алгоритм смог обнаружить среди всех объектов заданного класса;
- F-мера — среднее гармоническое точности и полноты.

Метрики являются стандартными для задач классификации.

2.2. Модели на основе лингвистических характеристик

Лингвистические модели подразумевают подсчёт стандартных статистических характеристик текста. Для исследования выбраны параметры, которые успешно показали себя в обработке англоязычных текстов [17]. Авторы рассмотрели уровень слов и символов, включающий в себя следующие значения:

- частота встречаемости каждой буквы как доля среди всех букв в тексте;
- частота встречаемости каждого знака пунктуации или спецсимвола как доля среди всех знаков пунктуации и спецсимволов в тексте;
- средняя длина предложения в символах;
- средняя длина предложения в словах;
- средняя длина слова в символах.

Как отдельный подход к представлению текста, были использованы характеристики уровня структуры — n-граммы частей речи, где $n = 1, 2, 3, 4$. Текст преобразовывался в последовательность частей речи, затем искали 10 самых популярных частей речи ($n = 1$) среди всех текстов, 10 самых популярных биграмм ($n = 2$) частей речи среди всех текстов и т. д. для $n = 3, 4$. Таким образом был составлен набор из сорока n-грамм. После этого для каждого текста были найдены частоты встречаемости выбранных n-грамм как количество появлений заданной n-граммы, делённое на количество появлений всех сорока n-грамм в этом тексте.

Каждый текст из корпуса моделировался как числовой вектор характеристик. Элементы в векторах размещались по заданным позициям, каждая позиция представляла отдельную характеристику, чтобы при сопоставлении векторов параметры на одних и тех же позициях соответствовали друг другу по смыслу. Вычисление характеристик происходило с использованием Python-библиотеки Stanza [18].

2.3. Большие языковые модели

Большие языковые модели (Large Language Model, LLM) — это программные средства для автоматизации обработки естественного текста, созданные на основе современных нейросетевых архитектур и предобученные на огромных корпусах данных. Для сравнения были выбраны варианты LLM BERT и GPT.

- Мультиязычный BERT и BERT для русского языка от DeepPavlov:
  - rubert-base-cased — базовая версия BERT для русского языка, учитывающая регистр;
  - rubert-base-cased-conversational — специализированная версия RuBERT, оптимизированная для работы с разговорной речью;
  - rubert-base-cased-sentence — версия RuBERT, оптимизированная для задач сравнения предложений;
  - bert-base-multilingual-cased-sentence — версия BERT, аналогичная предыдущей, но оптимизированная для кросс-язычной работы;
  - bert-base-bg-pl-ru-cased — базовая предобученная BERT-модель, сфокусированная на четырех славянских языках.
- Модели GPT от OpenAI:
  - openai-community/gpt2 — базовая версия GPT-2, содержит 124 миллиона параметров;
  - openai-community/gpt2-medium — средняя версия GPT-2 для задач, требующих большей связности, содержит 355 миллионов параметров;
  - openai-community/gpt2-large — большая версия GPT-2, используемая для задач, где важны качество и детализация, содержит 774 миллиона параметров;

  - openai-community/gpt2-xl — наибольшая версия GPT-2 для английского языка, используемая для задач с большим количеством контекстов, содержит более 1.5 миллиардов параметров.

Все модели были взяты из открытого ресурса Hugging Face[1]. Модели сопровождаются собственными токенизаторами, которые преобразовывают исходный текст на естественном языке в последовательность токенов, подающихся на вход модели. На выходе для каждого текста формировался эмбеддинг — вектор действительных чисел фиксированного размера.

2.4. Классификаторы машинного обучения

Для классификации по уровням CEFR многомерных характеристических векторов текстов использованы алгоритмы машинного обучения — математические алгоритмы, обученные на данных для решения конкретной задачи. Классификатор выявляет закономерности в обучающих данных и использует их для предсказаний на новых, ранее не встречавшихся. Авторы использовали описанные ниже классификаторы. Программная реализация взята из библиотеки scikit-learn.

- *k*-Nearest Neighbors (kNN): алгоритм классификации или регрессии, основанный на принципе близости векторов. Предсказание для нового объекта определяется «голосованием» (для классификации) или усреднением (для регрессии) значений k ближайших к нему точек в обучающей выборке. Параметры, использованные авторами:
  - n_neighbors=[3, 5, 7, 9];
  - weights=['uniform', 'distance'];
  - metric=['euclidean', 'cosine', 'manhattan'];
  - pca_n_components=[0.85, 0.90, 0.95].
- Support Vector Machines (SVC): алгоритм переводит исходные вектора в пространство более высокой размерности и ищет разделяющую гиперплоскость с максимальным расстояние между ней и опорными векторами каждого класса. Параметры, использованные авторами:
  - kernel=linear;
  - C=0.025;
  - random_state=42.
- AdaBoostClassifier: ансамбль деревьев решений, объединяющий прогнозы отдельных слабых моделей для получения окончательного предсказания. Параметры, использованные авторами:
  - n_estimators: 50;
  - random_state: 100.
- CatBoost: готовый алгоритм градиентного бустинга от Yandex, оптимизированный для работы с категориальными признаками. Автоматически обрабатывает категориальные переменные без предварительного кодирования, используя порядковое кодирование на основе статистики. Параметры, использованные авторами:
  - iterations: 1500;
  - learning_rate: 0.03;
  - depth: 8;
  - random_seed: 42;
  - l2_leaf_reg: 5.
- LightGBM (Light Gradient Boosting Machine): фреймворк градиентного бустинга от Microsoft, использующий древовидное обучение на основе листьев (leaf-wise growth), а не уровней (level-wise). Оптимизирован для скорости и низкого потребления памяти. Параметры, использованные авторами:
  - boosting_type: ['gbdt', 'dart', 'goss'];

[1] https://huggingface.co/models

Table 1. Texts distribution by levels **Таблица 1.** Распределение текстов по уровням

Уровень	Число текстов
A1	282
A2	367
B1	568
B2	252
C1	305
C2	130

  - num_leaves: [31, 63, 127, 255];
  - learning_rate: [0.01, 0.05, 0.1];
  - n_estimators: [100, 200, 300, 500];
  - subsample: [0.6, 0.8, 1.0];
  - colsample_bytree: [0.6, 0.8, 1.0];
  - reg_alpha: [0, 0.1, 0.5];
  - reg_lambda: [0, 0.1, 0.5];
  - min_child_samples: [5, 10, 20].
- XGBoost (eXtreme Gradient Boosting): алгоритм градиентного бустинга, сочетающий деревья решений с оптимизацией на основе вторых производных. Параметры, использованные авторами:
  - n_estimators: stats.randint(100, 500);
  - max_depth: stats.randint(3, 10);
  - learning_rate: stats.uniform(0.01, 0.3);
  - subsample: stats.uniform(0.6, 0.4);
  - colsample_bytree: stats.uniform(0.6, 0.4);
  - gamma: stats.uniform(0, 0.5);
  - reg_alpha: stats.uniform(0, 1);
  - reg_lambda: [0, 0.1, 0.5];
  - reg_lambda: stats.uniform(0, 1).
- StackingClassifier: ансамблевый метод, комбинирующий несколько базовых моделей (SVM, kNN, деревья) с помощью мета-классификатора. Предсказания базовых моделей становятся признаками для обучения финального классификатора. В качестве базовых моделей авторы использовали SVC, XGBoost и LightGBM классификаторы, в качестве мета-классификатора выступила логистическая регрессия. Для оптимизации гиперпараметров использовался фреймворк Optuna, параметры настройки соответствуют указанным в предыдущих пунктах.

Дополнительно в качестве классификатора использовалась модель многослойного персептрона из библиотеки scikit-learn: MLPClassifier с параметрами alpha=1, max_iter=1000, random_state=42.

2.5. Корпус текстов

Для проведения экспериментов был собран корпус русскоязычных текстов, состоящий из фрагментов статей, художественных произведений и учебных материалов, собранных из различных источников в сети Интернет. Все тексты размечены по шкале CEFR. Общий объем составляет 1904 экземпляра. В таблице 1 представлено распределение текстов по классам.

3. Результаты экспериментов

Для классификации в первой и второй группах экспериментов данные были разделены в пропорции 80 % обучающей выборки и 20 % тестовой.

Table 2. Linguistic models classification results for 6 levels A1–C2

Таблица 2. Результаты классификации лингвистических моделей для 6 уровней A1–C2

Модель	Классификатор	Точность	Полнота	F-мера
уровень слов и символов	AdaBoostClassifier	0.47	0.38	0.42
	KNN	0.45	0.42	0.44
	MLPClassifier	0.48	0.45	**0.47**
уровень структуры	AdaBoostClassifier	0.38	0.33	0.35
	SVC	0.40	0.38	0.39
	MLPClassifier	0.39	0.40	**0.40**

Table 3. Linguistic models classification results for 3 levels A, B, C

Таблица 3. Результаты классификации лингвистических моделей для 3 уровней A, B, C

Модель	Классификатор	Точность	Полнота	F-мера
уровень слов и символов	SVC	0.60	0.56	0.58
	AdaBoostClassifier	0.61	0.58	0.59
	MLPClassifier	0.64	0.63	**0.63**
уровень структуры	AdaBoostClassifier	0.55	0.53	0.54
	SVC	0.56	0.56	0.56
	MLPClassifier	0.59	0.59	**0.59**

Рассмотрим модели на основе лингвистических характеристик. В таблицах 2 и 3 приведены лучшие результаты определения шести и трёх уровней текстов соответственно.

В целом качество классификации невелико. Характеристики уровня слов и символов опережают структурные характеристики. Интересно, что в обоих случаях лучшим классификатором оказался многослойный персептрон. В таблице 4 приведена матрица ошибок лингвистической модели уровня слов и символов. Заметно, что основные ошибки происходят при разделении соседних уровней.

В таблице 5 представлены результаты работы алгоритмов, обученных на эмбедингах моделей BERT для для классификации на шесть категорий.

Лучших результатов удалось достичь на эмбеддингах rubert-base-cased-conversational, максимальное значение F-меры оказалось у ансамблевого классификатора StackingClassifier. Рассмотреть работу алгоритма более детально позволяет матрица ошибок, она представлена в таблице 6. Модель хорошо справляется с крайними уровнями A1 и C2. Почти отсутствуют ошибки между крупными группами A1–B1 и B2–C2. Главная проблема в классе B2: 36 % текстов B2 ошибочно понижены до B1, 20 % текстов B2 ошибочно повышены до C1.

В таблице 7 представлены результаты классификаторов, обученных на эмбеддингах GPT. Из неё видно, что эмбеддинги GPT решают задачу существенно хуже BERT. Модели gpt2-large, gpt2-xl и gpt2-medium мало отличаются друг от друга. В среднем лучших результатов удалось достичь классификаторам, обученным на эмбеддингах gpt2-medium и gpt2-large, а максимальное значение F-меры продемонстрировал алгоритм XGBoost. 39 % текстов уровня A2, 42 % текстов уровня B2 и 31 %

Table 4. MLPClassifier confusion matrix for 6 classes A1–C2

Таблица 4. Матрица ошибок MLPClassifier для 6 классов A1–C2

	A1	A2	B1	B2	C1	C2
A1	28	22	13	2	0	0
A2	13	41	24	6	2	1
B1	12	20	91	6	11	2
B2	6	12	19	13	17	5
C1	1	2	23	6	40	4
C2	0	2	7	2	7	16

Table 5. BERT embeddings classification results for 6 levels A1–C2

Таблица 5. Результаты классификации эмбеддингов BERT для 6 уровней A1–C2

Модель эмбеддингов	Классификатор	Точность	Полнота	F-мера
rubert-base-cased-sentence	CatBoost	0.44	0.44	0.43
	kNN	0.51	0.48	0.48
	LightGBM	0.52	0.50	0.49
	XGBoost	0.53	0.52	0.50
	StackingClassifier	0.53	0.51	0.51
bert-base-multilingual-cased-sentence	CatBoost	0.44	0.44	0.43
	kNN	0.50	0.45	0.45
	LightGBM	0.48	0.46	0.45
	XGBoost	0.49	0.50	0.49
	StackingClassifier	0.53	0.52	0.52
bert-base-bg-cs-pl-ru-cased	LightGBM	0.56	0.56	0.54
	CatBoost	0.59	0.58	0.58
	StackingClassifier	0.61	0.59	0.58
	kNN	0.62	0.56	0.58
	XGBoost	0.64	0.61	0.60
rubert-base-cased	CatBoost	0.53	0.54	0.53
	XGBoost	0.59	0.58	0.57
	kNN	0.61	0.61	0.60
	LightGBM	0.63	0.61	0.60
	StackingClassifier	0.66	0.66	0.66
rubert-base-cased-conversational	LightGBM	0.61	0.60	0.58
	CatBoost	0.60	0.61	0.60
	XGBoost	0.65	0.64	0.63
	kNN	0.67	0.62	0.63
	StackingClassifier	0.68	0.67	**0.67**

Table 6. StackingClassifier confusion matrix for 6 classes A1–C2

Таблица 6. Матрица ошибок модели StackingClassifier для 6 классов A1–C2

	A1	A2	B1	B2	C1	C2
A1	38	12	5	1	0	0
A2	2	43	29	0	0	0
B1	5	14	85	3	7	0
B2	0	3	18	18	10	1
C1	0	2	6	5	47	1
C2	0	0	0	0	3	23

текстов уровня C1 ошибочно определяются как B1. Как и для лексических моделей чаще всего путаются соседние уровни, например, 42 % текстов уровня B2 отнесены к B1, 22 % к C1.

Эксперименты с тремя классами подтвердили складывающуюся тенденцию для качества классификации двух типов эмбеддингов. В таблице 9 представлены лучшие результаты среди всех использованных моделей.

Максимального качества классификации удалось достичь на эмбеддингах модели разговорного русского языка RuBERT.

Table 7. GPT embeddings classification results for 6 levels A1–C2

Таблица 7. Результаты классификации эмбеддингов GPT для 6 уровней A1–C2

Модель эмбеддингов	Классификатор	Точность	Полнота	F-мера
gpt2	LightGBM	0.43	0.40	0.39
	kNN	0.43	0.42	0.41
	CatBoost	0.44	0.42	0.42
	XGBoost	0.48	0.45	0.44
	StackingClassifier	0.47	0.45	0.44
gpt2-large	LightGBM	0.49	0.49	0.48
	kNN	0.49	0.49	0.49
	XGBoost	0.52	0.51	0.50
	CatBoost	0.52	0.51	0.51
	StackingClassifier	0.52	0.52	0.51
gpt2-medium	CatBoost	0.50	0.48	0.46
	XGBoost	0.51	0.49	0.49
	kNN	0.51	0.50	0.50
	LightGBM	0.54	0.53	0.52
	StackingClassifier	0.53	0.52	0.52
gpt2-xl	kNN	0.43	0.43	0.43
	LightGBM	0.50	0.49	0.49
	CatBoost	0.51	0.51	0.50
	StackingClassifier	0.52	0.51	0.50
	XGBoost	0.56	0.53	**0.52**

Table 8. XGBoost confusion matrix for 6 classes A1–C2

Таблица 8. Матрица ошибок модели XGBoost для 6 классов A1–C2

	A1	A2	B1	B2	C1	C2
A1	26	17	13	0	0	0
A2	9	30	29	2	4	0
B1	7	14	82	4	6	1
B2	2	3	21	13	11	0
C1	1	4	19	0	37	0
C2	0	0	6	0	5	15

Table 9. Best classification results for 3 levels A, B, C

Таблица 9. Лучшие результаты классификации для 3 уровней A, B, C

Модель эмбеддингов	Классификатор	Точность	Полнота	F-мера
gpt2-large	kNN	0.65	0.65	0.65
	XGBoost	0.67	0.66	0.66
	LightGBM	0.69	0.67	0.67
	CatBoost	0.68	0.68	0.68
	StackingClassifier	0.72	0.71	0.71
rubert-base-cased-conversational	CatBoost	0.73	0.73	0.72
	kNN	0.74	0.73	0.74
	LightGBM	0.77	0.76	0.76
	XGBoost	0.77	0.76	0.76
	StackingClassifier	0.78	0.77	**0.77**

Заключение

Результаты сравнения двух видов моделей для классификации русскоязычных текстов по уровням CEFR показали преимущество эмбеддингов LLM. Максимальное значение F-меры для прогнозирования шести категорий текста равно 0.67 и достигнуто с использование модели rubert-base-cased-conversational. Интересно, что он получен без дополнительной адаптации языковой модели. Также стоит отметить, что качество определения уровня больше зависит от модели, чем от алгоритма классификации машинного обучения. Результаты многих алгоритмов отличаются друг от друга не более чем на 0.01–0.02, особенно это касается ансамблевых методов.

Эксперименты с классификацией на шесть категорий CEFR для русского языка очень похожи на результаты исследований англоязычных текстов из корпуса CEFR Levelled English Texts [8] с F-мерой равной 0.69, лучшее значение для русскоязычного корпуса — 0.67. Корпус CEFR Levelled English сопоставим по объёму с собранным авторами: 1497 и 1904 текста соответственно. Однако классификация на три уровня лучше для английского языка (0.90), а матрица ошибок показывает большую сложность разделения внутри отдельного уровня на подуровни. Для русского языка характерны ошибки между всеми соседними категориями, интересно что такая же картина наблюдается для другого англоязычного корпуса BEA-2019 большего объёма: 3300 текстов. Похожие результаты с эмбеддингами BERT для шести категорий CEFR описаны для французского (F-мера равна 0.57) и шведского (F-мера равна 0.74) [9]. Объём французского корпуса текстов — 6569 существенно больше шведского — 90. Возможно, в меньшем корпусе тексты разных категорий сильнее отличаются друг от друга, в корпусе большого размера это сделать сложнее. Классификация на три класса ожидаемо оказалась намного успешнее, чем на шесть: F-мера равна 0.77. Это меньше, чем для английского, но близко к результатам в других национальных языках, например, для большого корпуса из 22820 арабских текстов модель ArabicBERT показала F-меру равную 0.80 [11].

Повысить качество классификации моделей может сбор и разметка корпусов большего объёма, но это является трудоёмкой задачей. Вычисление сложных лингвистических характеристик, отражающих свойства естественного языка, также может способствовать улучшению решения, особенно в области интерпретации результата и построения обратной связи с обучающимися.

References

[1] N. V. Bordovskaya, E. A. Koshkina, M. A. Tikhomirova, and L. A. Melkaya, “Blended educational technologies in higher education: Systematic review of domestic publications”, *Vysshee obrazovanie v Rossii = Higher Education in Russia*, vol. 31, no. 8-9, pp. 58–78, 2022, in Russian. DOI: 10.31992/0869-3617-2022-31-8-9-58-78.

[2] A. W. Zaki and R. Darmi, “CEFR: Education towards 21st century of learning. Why matters”, *Journal of Social Science and Humanities*, vol. 4, no. 2, pp. 14–20, 2021. DOI: 10.26666/rmp.jssh.2021.2.3.

[3] T. Gaillat *et al.*, “Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach”, *ReCALL*, vol. 34, no. 2, pp. 130–146, 2022. DOI: 10.1017/S095834402100029X.

[4] I. Natova, “Estimating CEFR reading comprehension text complexity”, *The Language Learning Journal*, vol. 49, no. 6, pp. 699–710, 2021. DOI: 10.1080/09571736.2019.1665088.

[5] G. Fakhretdinova, L. M. Zinnatullina, F. T. Galeeva, and E. Valeeva, “The CEFR for languages: Research perspectives in foreign language teaching in engineering university”, in *International Conference on Interactive Collaborative Learning*, Springer, 2021, pp. 225–232. DOI: 10.1007/978-3-030-93907-6_24.

[6] A. Dmitrieva, A. Laposhina, and M. Lebedeva, “A comparative study of educational texts for native, foreign, and bilingual young speakers of russian: Are simplified texts equally simple?”, *Frontiers in Psychology*, vol. 12, p. 703 690, 2021. DOI: 10.3389/fpsyg.2021.703690.

[7] V. J. Schmalz and A. Brutti, "Automatic assessment of english CEFR levels using BERT embeddings", in *Computational Linguistics CliC-it 2021*, 2022, pp. 293–299.

[8] N. Lagutina, K. Lagutina, A. Brederman, and N. Kasatkina, "Text classification by CEFR levels using machine learning methods and the BERT language model", *Automatic Control and Computer Sciences*, vol. 58, no. 7, pp. 869–878, 2024. DOI: 10.3103/S0146411624700329.

[9] R. M. Sánchez, D. Alfter, S. Dobnik, M. Szawerna, and E. Volodina, "Jingle BERT, jingle BERT, frozen all the way: Freezing layers to identify CEFR levels of second language learners using BERT", in *Swedish Language Technology Conference and NLP4CALL*, 2024, pp. 137–152. DOI: 10.3384/ecp211011.

[10] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? an analysis of BERT's attention", in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019, pp. 276–286. DOI: 10.18653/v1/W19-4828.

[11] N. Khallaf and S. Sharoff, "Automatic difficulty classification of arabic sentences", in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 105–114.

[12] D. Kogan *et al.*, *Ace-CEFR–a dataset for automated evaluation of the linguistic difficulty of conversational texts for LLM applications*, 2025. arXiv: 2506.14046 [cs.CL].

[13] H. Ma, J. Wang, and L. He, "Linguistic features distinguishing students' writing ability aligned with CEFR levels", *Applied Linguistics*, vol. 45, no. 4, pp. 637–657, 2024. DOI: 10.1093/applin/amad054.

[14] V. Franzoni, G. Biondi, A. Milani, *et al.*, "Morpho-phraseological based classification of CEFR Italian L2 learner writing proficiency", *IEEE ACCESS*, vol. 12, pp. 156 433–156 441, 2024. DOI: 10.1109/ACCESS.2024.3485988.

[15] A. N. Laposhina and M. Y. Lebedeva, "Textometr: An online tool for automated complexity level assessment of texts for Russian language learners", *Russian language studies*, vol. 19, no. 3, pp. 331–345, 2021, in Russian. DOI: 10.22363/2618-8163-2021-19-3-331-345.

[16] I. Glišić, C. L. Richter, and A. K. Ingason, "Testing relevant linguistic features in automatic CEFR skill level classification for Icelandic", in *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, 2025, pp. 217–222.

[17] N. Lagutina, K. Lagutina, A. Brederman, and N. Kasatkina, "Text classification by CEFR levels using machine learning methods and BERT language model", *Modeling and Analysis of Information Systems*, vol. 30, no. 3, pp. 202–213, 2023, in Russian. DOI: 10.18255/1818-1015-2023-3-202-213.

[18] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 101–108. DOI: 10.18653/v1/2020.acl-demos.14.