MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 30, NO. 3, 2023
journal homepage: www.mais-journal.ru

THEORY OF DATA

# Text classification by CEFR levels using machine learning methods and BERT language model

N. S. Lagutina[1], K. V. Lagutina[1], A. M. Brederman[1], N. N. Kasatkina[1]

DOI: 10.18255/1818-1015-2023-3-202-213

[1]P.G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 93A30, 68Q60
Research article
Full text in Russian

Received August 14, 2023
After revision August 25, 2023
Accepted August 30, 2023

This paper presents a study of the problem of automatic classification of short coherent texts (essays) in English according to the levels of the international CEFR scale. Determining the level of text in natural language is an important component of assessing students knowledge, including checking open tasks in e-learning systems. To solve this problem, vector text models were considered based on stylometric numerical features of the character, word, sentence structure levels. The classification of the obtained vectors was carried out by standard machine learning classifiers. The article presents the results of the three most successful ones: Support Vector Classifier, Stochastic Gradient Descent Classifier, LogisticRegression. Precision, recall and F-score served as quality measures. Two open text corpora, CEFR Levelled English Texts and BEA-2019, were chosen for the experiments. The best classification results for six CEFR levels and sublevels from A1 to C2 were shown by the Support Vector Classifier with F-score 67 % for the CEFR Levelled English Texts. This approach was compared with the application of the BERT language model (six different variants). The best model, bert-base-cased, provided the F-score value of 69 %. The analysis of classification errors showed that most of them are between neighboring levels, which is quite understandable from the point of view of the domain. In addition, the quality of classification strongly depended on the text corpus, that demonstrated a significant difference in F-scores during application of the same text models for different corpora. In general, the obtained results showed the effectiveness of automatic text level detection and the possibility of its practical application.

**Keywords:** natural language processing; text classification; CEFR; BERT

INFORMATION ABOUT THE AUTHORS

| | |
|---|---|
| Nadezhda S. Lagutina | orcid.org/0000-0002-6137-8643. E-mail: lagutinans@gmail.com<br>PhD, associate professor. |
| Ksenia V. Lagutina<br>corresponding author | orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru<br>PhD, associate professor. |
| Anastasya M. Brederman | orcid.org/0009-0003-1741-0604. E-mail: anastasyabrederman@mail.ru<br>student. |
| Natalia N. Kasatkina | orcid.org/0000-0002-6757-9622. E-mail: ninet75@mail.ru<br>PhD, associate professor. |

**Funding:** This study was supported by YarSU Development Program until 2030, project No. GM-2023-123061600058-4 “Development of an automated system for the development of mediative competence in language education”.

**For citation**: N. S. Lagutina, K. V. Lagutina, A. M. Brederman, and N. 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.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 30, № 3, 2023
сайт журнала: www.mais-journal.ru

THEORY OF DATA

# Классификация текстов по уровням CEFR с использованием методов машинного обучения и языковой модели BERT

Н. С. Лагутина[1], К. В. Лагутина[1], А. М. Бредерман[1], Н. Н. Касаткина[1]

DOI: 10.18255/1818-1015-2023-3-202-213

[1]Ярославский государственный университет им. П.Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 14 августа 2023 г.
После доработки 25 августа 2023 г.
Принята к публикации 30 августа 2023 г.

В данной работе представлено исследование задачи автоматической классификации коротких связных текстов (эссе) на английском языке по уровням международной шкалы CEFR. Определение уровня текста на естественном языке является важной составляющей оценки знаний учащихся, в том числе для проверки открытых заданий в системах электронного обучения. Для решения этой задачи были рассмотрены векторные модели текста на основе стилометрических числовых характеристик уровня символов, слов, структуры предложения. Классификация полученных векторов осуществлялась стандартными классификаторами машинного обучения. В статье приведены результаты трёх наиболее успешных: Support Vector Classifier, Stochastic Gradient Descent Classifier, LogisticRegression. Оценкой качества послужили точность, полнота и F-мера. Для экспериментов были выбраны два открытых корпуса текстов CEFR Levelled English Texts и BEA-2019. Лучшие результаты классификации по шести уровням и подуровням CEFR от A1 до C2 показал Support Vector Classifier с F-мерой 67 % для корпуса CEFR Levelled English Texts. Этот подход сравнивался с применением языковой модели BERT (шесть различных вариантов). Лучшая модель bert-base-cased обеспечила значение F-меры 69 %. Анализ ошибок классификации показал, что большая их часть допущена между соседними уровнями, что вполне объяснимо с точки зрения предметной области. Кроме того, качество классификации сильно зависело от корпуса текстов, что продемонстрировало существенное различие F-меры в ходе применения одинаковых моделей текста для разных корпусов. В целом, полученные результаты показали эффективность автоматического определения уровня текста и возможность его практического применения.

**Ключевые слова:** автоматическая обработка текста; классификация текста; CEFR; BERT

ИНФОРМАЦИЯ ОБ АВТОРАХ

| | |
|---|---|
| Надежда Станиславовна Лагутина | orcid.org/0000-0002-6137-8643. E-mail: lagutinans@gmail.com<br>канд. физ.-мат. наук, доцент. |
| Ксения Владимировна Лагутина<br>автор для корреспонденции | orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru<br>канд. тех. наук, доцент. |
| Анастасия Михайловна Бредерман | orcid.org/0009-0003-1741-0604. E-mail: anastasyabrederman@mail.ru<br>студент. |
| Наталья Николаевна Касаткина | orcid.org/0000-0002-6757-9622. E-mail: ninet75@mail.ru<br>канд. пед. наук, доцент. |

**Финансирование:** Исследование выполнено за счет средств Программы развития ЯрГУ до 2030 года, проект № GM-2023-123061600058-4 «Разработка автоматизированной системы развития медиативной компетенции в языковом образовании».

**Для цитирования**: N. S. Lagutina, K. V. Lagutina, A. M. Brederman, and N. 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.

## Введение

Автоматизированная оценка эссе (automated essay scoring, AES) — это способ моделирования работы человека-эксперта в области языкознания и педагогики. Определение качества текста на естественном языке является важной составляющей оценки знаний учащихся, в том числе для проверки открытых заданий в системах электронного обучения [1, 2]. Кроме того, текст является одной из основ для коммуникации людей, что порождает необходимость выявления сложности текстов, их качества и возможности понимания целевой аудиторией [3].

Развитие систем AES началось с 60-х годов прошлого века и опиралось в первую очередь на правила грамматики. По мере развития информационных технологий в области компьютерной лингвистики исследователи добавляли анализ стиля текста, его структуры, связности на основе параметров текста различной степени сложности. Большинство работ в этой области используют статистические признаки, такие как функции «мешка слов» (Bag of Words, BoW), количество предложений и т. п. Однако качество текста во многом определяется его связностью, анализом семантики, но даже использование контекстных параметров типа word2vec не решает эту проблему и не повышает качество AES до должного уровня [4]. Новые возможности предоставляет развитие современных языковых моделей [5], в частности модель BERT [6]. Обзор современных работ показывает, что AES растущая область исследований с большим набором потенциально применимых методов, но все еще не зрелая, особенно в сфере практического применения [1].

Отдельным вопросом AES является способ и шкала оценки текстов. Исследователи используют системы на основе подсчета баллов [7], формулируют наборы проверяемых критериев [8], применяют существующие стандарты, описывающие уровни владения языком (общеевропейские компетенции владения иностранным языком CEFR, стандарты преподавания иностранных языков ACTFL, канадские критерии оценки уровня языка CLB, межведомственный круглый стол по вопросам языковой подготовки ILR). Значительная часть работ в образовательной сфере и области AES делит эссе по уровням CEFR [5, 9].

Авторы данной статьи рассмотрели задачу AES как задачу классификации текстов по трём уровням CEFR (A — начальный, B — средний и C — высокий) и шести подуровням (A1, A2, B1, B2, C1, C2) и поставили цель систематизировать и проанализировать результаты классификации с использованием стилометрических параметров текста и стандартных классификаторов машинного обучения и сравнить их с результатами применения языковых моделей BERT.

## 1. Аналогичные работы

Методы классификации и анализа текстов на естественном языке бурно развиваются в последнее десятилетие, поэтому при выборе инструментов для исследования и сравнения результатов авторы статьи сосредоточились на достижениях последних лет.

Авторы работы [10] использовали модель bert-base-uncased для классификации Кембриджской базы данных открытого языка (EFCAMDAT) и Кембриджского учебного корпуса по английскому языку (CLC-FCE) на пять уровней шкалы CEFR: от A1 до C1. Точность классификации (accuracy) достигла 61.7 ± 1.8 %. Исследователи отметили, что качество сильно зависит от размера корпуса для обучения.

В статье [11] учёные классифицировали по уровням CEFR отдельные предложения. Для сбора и разметки собственного корпуса из 17 000 английских предложений были привлечены высококвалифицированные преподаватели, что обеспечило качество данных. Каждому предложению ставился в соответствие числовой вектор на основе частот слов уровней CEFR, классификация по шести уровням осуществлялась с помощью косинусного сходства между векторами. Предложенный подход показал F-меру со средним значением 84.5 ± 0.7 %. Авторы сравнили этот результат с моделью

BERT, F-мера оказалась 82.5 ± 0.9 % и моделью «мешок слов» с классификатором SVM, где F-мера чуть больше 52 %.

Зависимость качества классификации от качества используемого корпуса данных отмечается в работе [12]. Авторы указывают на разнородность и несбалансированность текстов, собираемых среди изучающих английский язык как второй. Для преодоления этих проблем для моделирования текстов они предложили использовать набор параметров, ориентированных на уровень владения языком, а не на текстовые функции. Эксперименты были проведены в Международной корпусной сети азиатских изучающих английский язык (ICNALE) на наборе данных, включающем эссе от 2 800 авторов. Исследователи провели классификацию по уровням CEFR с помощью многослойного персептрона и линейной регрессии и получили среднюю F-меру 63 % и 43 % соответственно. Аналогичный эксперимент с корпусом EFCAMDAT показал значение F-меры 96 % и 83 %.

Параметры на основе частоты встречаемости различных языковых конструкций были использованы для моделирования текста в исследовании [13]. Эти конструкции, предложенные экспертами преподавателями и лингвистами, авторы назвали микросистемами. Для экспериментов использовались корпуса EFCAMDAT и CEFR-ASAG. Классификация полиномиальной логистической регрессией на шесть уровней CEFR показала среднюю точность (accuracy) 75 % и 95 % при отделении уровня начинающих A от продвинутых B, что может быть полезно для автоматического разделения учащихся на группы.

57 числовых характеристик использовались в статье [14]. Они разделены на четыре группы: меры синтаксической сложности, на основе анализа синтаксических деревьев; меры лексического богатства; $n$-граммы слов, для $n$ от двух до пяти; колмогоровская сложность как мера количества информации в строке. Классификатором служила рекуррентная нейронная сеть, корпус данных EFCAMDAT. Качество классификации оказалось выше для начального и среднего уровней владения языком CEFR (от A1 до B2) с F-мерой в диапазоне от 73 % до 81 % по сравнению с уровнями C1 и C2, где F-мера упала до 61 % для уровня C1 и 42 % для C2. Матрица ошибок показала, что их большинство возникло в основном в смежных категориях.

Таким образом, для классификации уровней CEFR применяются разнообразные лексические параметры, отражающие специфику предметной области. Однако, авторы данной статьи не увидели системных исследований стилометрических характеристик текста для этой цели, хотя они вполне успешно используются для AES в целом [15]. Кроме того, очень перспективным методом решения задачи выглядит работа с разными моделями BERT. В пользу этого говорят результаты успешного применения BERT для задач оценки эссе [16] и сложности текста [17].

## 2. Метод автоматического определения уровня владения языком

### 2.1. Корпуса текстов

Для исследования уровня владения английским языком были взяты два открытых корпуса, тексты в которых размечены по международному стандарту CEFR.

Корпус CEFR Levelled English Texts содержит 1 494 текста из открытых источников The British Council, ESLFast и корпуса cnn-dailymail. Он опубликован на Kaggle (https://www.kaggle.com/datasets/amontgomerie/cefr-levelled-english-texts).

Второй корпус появился в рамках соревнования BEA-2019 [18], посвящённого поиску грамматических ошибок. Открытая часть корпуса содержит 3 350 текстов, 3 300 из которых размечены по шкале CEFR. Тексты были агрегированы из корпусов Cambridge English Write & Improve и LOCNESS.

Дополнительно для экспериментов корпуса объединялись в один набор текстов. Статистические данные о категориях текстов в корпусах представлены в таблице 1.

Оба корпуса имеют разные принципы и источники сборки, поэтому в качестве первичной обработки данных были рассчитаны базовые статистические параметры: среднее, модальное, ме-

**Table 1.** Number of texts of different CEFR levels in corpora

**Таблица 1.** Количество текстов различных уровней CEFR в корпусах

| Корпус | Всего | A | A1 | A2 | B | B1 | B2 | C | C1 | C2 |
|---|---|---|---|---|---|---|---|---|---|---|
| CEFR Levelled English Texts | 1 494 | 560 | 288 | 272 | 491 | 205 | 286 | 443 | 241 | 202 |
| BEA-2019 | 3 300 | 1 430 | 585 | 845 | 1 100 | 631 | 469 | 770 | 483 | 287 |
| Объединение корпусов | 4 794 | 1 990 | 873 | 1 117 | 1 591 | 836 | 755 | 1 213 | 724 | 489 |

**Table 2.** Statistical indicators of the numbers of words in corpora

**Таблица 2.** Статистические показатели количества слов в корпусах

| Корпус | Уровень | Среднее значение | Мода | Медиана | Макс. | Мин. |
|---|---|---|---|---|---|---|
| CEFR Levelled English Texts | **Всего** | **424.82** | **125** | **313** | **2 292** | **37** |
| CEFR Levelled English Texts | A1 | 97.94 | 123 | 106 | 307 | 37 |
| CEFR Levelled English Texts | A2 | 232.63 | 107 | 291 | 1 222 | 70 |
| CEFR Levelled English Texts | B1 | 415.34 | 309 | 310 | 1 628 | 100 |
| CEFR Levelled English Texts | B2 | 509.83 | 299 | 342 | 2 668 | 97 |
| CEFR Levelled English Texts | C1 | 701.42 | 610 | 706 | 1 669 | 153 |
| CEFR Levelled English Texts | C2 | 708.93 | 518 | 678 | 2 292 | 115 |
| BEA-2019 | **Всего** | **188.09** | **194** | **176** | **1 612** | **31** |
| BEA-2019 | A1 | 85.00 | 107 | 78 | 364 | 31 |
| BEA-2019 | A2 | 155.76 | 155 | 148 | 686 | 33 |
| BEA-2019 | B1 | 203.64 | 202 | 188 | 1 046 | 46 |
| BEA-2019 | B2 | 226.58 | 151 | 199 | 1 612 | 73 |
| BEA-2019 | C1 | 262.50 | 199 | 228 | 1 407 | 44 |
| BEA-2019 | C2 | 271.09 | 209 | 229 | 1 093 | 47 |

дианное, максимальное и минимальное количества слов (таблица 2), и количество предложений в корпусах (таблица 3).

На основе вышеуказанных данных можно отметить, что показатели корпуса BEA-2019, во-первых, в среднем выше соответствующих показателей корпуса CEFR Levelled English Texts, а во-вторых, характеризуются меньшей вариацией одноимённых индикаторов. В дальнейшем эта закономерность могла стать фактором влияния на результаты классификации.

### 2.2. Модели текста

Тексты моделировались как вектора числовых характеристик на основе нескольких моделей:

- характеристики уровня символов;
- характеристики уровня слов;
- характеристики уровня структуры;
- эмбеддинги на основе BERT.

Первые три модели основаны на статистических и лингвистических характеристиках текстов. Алгоритмы для их вычисления представлены в предыдущих работах авторов [19, 20].

В группу *характеристик уровня символов* вошли следующие шесть стилометрических характеристик:

**Table 3.** Number of sentences in texts of different CEFR levels in corpora

**Таблица 3.** Количество предложений в текстах различных уровней CEFR в корпусах

| Корпус | Всего | A1 | A2 | B1 | B2 | C1 | C2 |
|---|---|---|---|---|---|---|---|
| CEFR Levelled English Texts | 36 235 | 4 572 | 6 023 | 5 381 | 7 552 | 7 344 | 5 363 |
| BEA-2019 | 35 171 | 3 109 | 7 366 | 7 288 | 6 050 | 7 044 | 4 314 |
| Объединение корпусов | 71 406 | 7 681 | 13 389 | 12 669 | 13 602 | 14 388 | 9 677 |

- общее количество символов в тексте;
- частоты букв латинского алфавита;
- частоты появления самых часто встречающихся знаков препинания (точка, занятая, двоеточие, точка с занятой, вопросительный знак, восклицательный знак, кавычки, скобки, тире);
- отношение общего количества букв латинского алфавита к общему количеству символов;
- отношение общего количества знаков препинания без учета точки к общему количеству символов;
- отношение общего количества цифр к общему числу символов.

Группа *характеристик уровня слов* была сформирована из 17 характеристик:

- количество слов в тексте;
- средняя длина слов в тексте;
- лексическое разнообразие — отношение числа разных лексем к общему числу слов;
- автосемантичность — отношение числа значащих слов (кроме служебных слов и местоимений) к общему числу слов;
- аналитичность — отношение числа служебных слов к общему числу слов;
- негация — отношение суммы отрицательных частиц к общему числу слов;
- субстантивность — отношение числа существительных к общему числу слов;
- отношение числа существительных во множественном числе к общему числу слов;
- глагольность — отношение числа всех глагольных форм к общему числу слов;
- отношение числа глаголов в прошедшем времени к общему числу слов;
- отношение числа глаголов в настоящем времени к общему числу слов;
- отношение числа модальных глаголов к общему числу слов;
- местоимённость — отношение местоимённых слов к общему числу слов;
- отношение числа личных местоимений к общему числу слов;
- адъектинность — отношение числа прилагательных к общему числу слов;
- отношение числа наречий к общему числу слов;
- отношение числа прилагательных и наречий в сравнительной форме к общему числу слов.

В группу *характеристик уровня структуры* документа вошли ещё четыре характеристики:

- количество предложений в тексте;
- средняя длина предложений;
- количество абзацев в тексте;
- частоты появления предложений, состоящих из определённого числа слов (от 1 до 39, 40-й признак включает в себя все остальные предложения).

Также вектора характеристик данных уровней конкатенировались, чтобы получить дополнительные модели-комбинации двух и трёх уровней.

*Эмбеддинги на основе BERT* строились на основе следующих моделей, открыто опубликованных на https://huggingface.co/:

- bert-base-cased [21] — классическая языковая модель BERT для английского языка, учитывающая регистр;
- bert-base-uncased — классическая языковая модель BERT для английского языка, не учитывающая регистр;
- bert-large-cased — увеличенная языковая модель BERT для английского языка, учитывающая регистр;
- distilbert-base-cased [22] — языковая модель, обученная на классической модели BERT и имеющая меньший размер;
- DeepPavlov/bert-base-cased-conversational — языковая модель на основе классической модели BERT, дообученная на текстах из социальных сетей и блогах;

- Intel/bert-base-uncased-mrpc — языковая модель на основе классической модели BERT, дообученная на корпусе GLUE MRPC.

BERT-эмбеддинги были выбраны, поскольку именно эта языковая модель в настоящее время достигает лучших результатов в различных задачах обработки текстов на естественном языке. Первые четыре модели из списка выше представляют собой вариации классического BERT, а пятая и шестая — BERT, дообученный на современных англоязычных интернет-текстах.

Таким образом, в исследовании сравнивались несколько векторных моделей текста: стилометрические трёх уровней вместе с их комбинациями и модели на основе шести типов эмбеддингов. Каждый текст представлялся как вектор числовых характеристик, вычисленных на основе одной из моделей.

### 2.3. Классификация текстов по уровню

После моделирования корпус текстов представлял собой матрицу векторов числовых характеристик. Каждому тексту (вектору) сопоставлялась категория, соответствующая одному из уровней владения языком. Классификация этих данных проводилась с помощью обучаемых методов двух типов: классификаторы машинного обучения и нейросетевые классификаторы.

Перед классификацией корпус текстов был разделён на обучающую, валидационную и тестовую выборки в пропорции 80 %/10 %/10 %. Валидационная выборка использовалась для подбора гиперпараметров.

Классификация векторов стилометрических характеристик проводилась на основе библиотеки машинного обучения scikit-learn (https://scikit-learn.org/stable/index.html). Базовые алгоритмы машинного обучения опираются исключительно на предоставленные данные из корпусов, из-за чего могут снабжать противоречивыми показателями и не давать стопроцентной точности предсказаний. Вместе с тем наиболее успешными в работе стали:

- SVC (Support Vector Classifier): классификатор, функционирующий на основе метода опорных векторов. Среди гиперпараметров выбраны: параметр регуляции — 1, параметр ядра — 0.1, функция ядра для преобразования пространства абзацев — линейное ядро (осуществляется линейная комбинация признаков).
- SGDClassifier (Stochastic Gradient Descent Classifier): классификатор, основанный на стохастическом градиентном спуске. Гиперпараметры: начальное состояние генератора случайных чисел — 42, параметр регуляции — 0.001, коэффициент смешивания регуляторов L1 и L2 — 0.75, максимальное число итераций градиентного спуска — 10 000.
- LogisticRegression: классификатор, который использует логистическую функцию для моделирования вероятности принадлежности точки к определенному классу. Базовыми аргументами стали: обратный коэффициент регуляции — 1, максимальное количество итераций — 100, функция регуляции — гребневая регрессия, алгоритм оптимизации — метод Ньютона-Кона.

Для классификации результатов BERT-моделей использовался однослойный перцептрон, объединённый в общую нейронную сеть с трансформером для языковой модели. Гиперпараметры были выбраны следующие: функция активации в выходном слое — линейная, оптимизатор — Adam, размер батча — 5.

Предсказания классификаторов оценивались при помощи общепринятых метрик: точности, полноты и F-меры.

## 3. Эксперименты

Тексты классифицировались несколькими способами: на все шесть уровней CEFR (A1, A2, B1, B2, C1, C2), на три уровня CEFR (A, B и C) и на пары подуровней A1 и A2, B1 и B2, C1 и C2. Первые два варианта позволяют оценить эффективность моделей в решении основной задачи исследования, а третий нужен для того, чтобы проанализировать особенности распознавания отдельных уровней.

**Table 4.** Text classification by 6 levels using machine learning

**Таблица 4.** Классификация текстов на 6 уровней с помощью машинного обучения

| Корпус | Классификатор | Точность | Полнота | F-мера |
|---|---|---|---|---|
| CEFR Levelled English Texts | Support Vector Classifier | **68** | **68** | **67** |
| CEFR Levelled English Texts | Stochastic Gradient Descent Classifier | 64 | 66 | 64 |
| CEFR Levelled English Texts | Logistic Regression | 68 | 68 | 67 |
| BEA-2019 | Support Vector Classifier | 40 | 40 | 39 |
| BEA-2019 | Stochastic Gradient Descent Classifier | 38 | 40 | 37 |
| BEA-2019 | Logistic Regression | **40** | **41** | **39** |
| Объединение корпусов | Support Vector Classifier | 47 | 47 | 46 |
| Объединение корпусов | Stochastic Gradient Descent Classifier | 42 | 45 | 42 |
| Объединение корпусов | Logistic Regression | 46 | 46 | 45 |

**Table 5.** Text classification by 6 levels with BERT

**Таблица 5.** Классификация текстов на 6 уровней с BERT

| Корпус | Модель | Точность | Полнота | F-мера |
|---|---|---|---|---|
| CEFR Levelled English Texts | bert-base-cased | **69** | **69** | **69** |
| CEFR Levelled English Texts | bert-base-uncased | 67 | 64 | 64 |
| CEFR Levelled English Texts | bert-large-cased | 68 | 66 | 67 |
| CEFR Levelled English Texts | distilbert-base-cased | 65 | 62 | 62 |
| CEFR Levelled English Texts | bert-base-cased-conversational | 69 | 68 | 68 |
| CEFR Levelled English Texts | Intel/bert-base-uncased-mrpc | 63 | 61 | 61 |
| BEA-2019 | bert-base-cased | 47 | 42 | 42 |
| BEA-2019 | bert-large-cased | **51** | **48** | **49** |
| BEA-2019 | bert-base-cased-conversational | 49 | 47 | 47 |
| Объединение корпусов | bert-base-cased | 51 | 50 | 50 |
| Объединение корпусов | bert-base-cased-conversational | 46 | 47 | 46 |

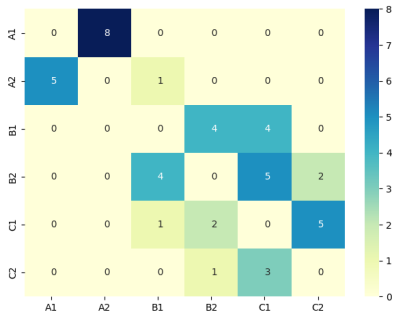
Результаты классификации текстов на шесть уровней владения языком с помощью алгоритмов машинного обучения получились невысокими. Результаты экспериментов приведены в таблице 4. Наиболее успешно с работой справился классификатор логистической регрессии (LogisticRegression). При этом классификация текстов из корпуса CEFR Levelled English Texts проходила на порядок успешнее классификации смежных корпусов: 67 % F-меры по сравнению с 39 и 45 % у корпусов BEA-2019 и объединённого соответственно.

Большую роль в успехе работы алгоритмов играли принципы векторизации данных: чем больше нюансов языка и лингвистических особенностей учитывалось при расчёте числовых параметров, тем выше были итоги классификации. Подобная закономерность указывает на наличие в языке неявных, но формализуемых критериев.

Расширенная классификация текстов при помощи BERT даёт показатели на 1–3 % выше показателей для алгоритмов машинного обучения. Результаты экспериментов BERT приведены в таблице 5. Лучше всего разделяются на категории тексты из корпуса CEFR Levelled English Texts при помощи классической модели BERT: 69 % F-меры. Использование вариаций данной языковой модели, в том числе большей по размеру и дообученных версий не даёт улучшения качества.

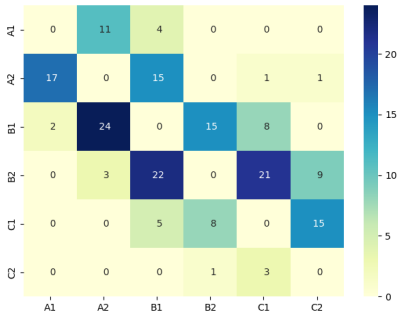
Классификация текстов BEA-2019 выполняется с F-мерой не более 49 %, объединённый корпус также классифицируется в лучшем случае с метриками около 50 %. В таблице 5 приведены BERT-модели, дающие лучшие результаты экспериментов.

На рис. 1 представлены матрицы ошибок для классификации обоих корпусов с помощью bert-base-cased. На рис. 2 представлены соответствующие матрицы ошибок для классификации с по-



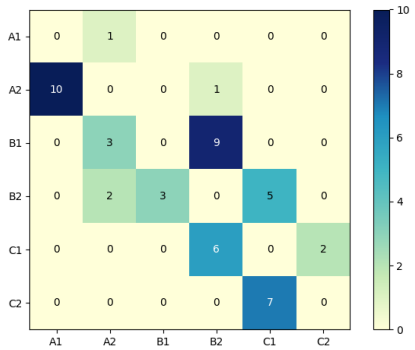

a)

**Fig. 1.** Number of errors of the model bert-base-cased for the corpus a) CEFR Levelled English Texts, b) BEA-2019



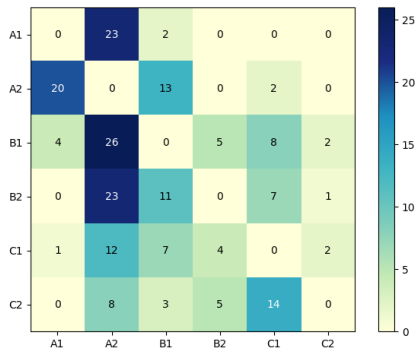

b)

**Рис. 1.** Количество ошибок модели bert-base-cased для корпуса a) CEFR Levelled English Texts, b) BEA-2019




a)

**Fig. 2.** Number of errors of the classifier Logistic Regression for the corpus a) CEFR Levelled English Texts, b) BEA-2019




b)

**Рис. 2.** Количество ошибок классификатора LogisticRegression для корпуса a) CEFR Levelled English Texts, b) BEA-2019

мощью метода опорных векторов. По вертикали указаны корректные уровни владения языком для текстов, а по горизонтали — уровни, за которые эти тексты были приняты ошибочно. На главных диагоналях условно отмечены нули. Видна общая закономерность: классификатор чаще путает между собой соседние уровни, а тексты, сильно отличающиеся по уровню владения языком, отделяются друг от друга хорошо.

При работе с методом опорных векторов меньше всего ошибок оказалось для подуровня A1. Чаще всего с ним путается подуровень A2. В то время как при работе с BERT меньше всего ошибок оказалось для подуровня C2. Подуровень C1 путается чаще всего с C2. Соответствующие матрицы ошибок у рассматриваемых классификаторов оказались обратно симметричны.

Подчеркивает последнее утверждение и закономерность исключений в корпусе CEFR Levelled English Texts: для BERT подуровни A смешиваются с соседними реже, чем подуровни C; для классификаторов, основанных на стилометрических характеристиках, подуровни A смешиваются с соседними чаще, чем подуровни C.

В корпусе CEFR Levelled English Texts ошибки сосредоточены в первую очередь в классификации пар подуровней. Подуровни A1 и A2 смешиваются практически только между собой, C1 и C2 чаще принимаются друг за друга, чем за другие подуровни. Подуровни B1 и B2 тоже путаются между собой, но как и C1, они также часто ошибочно классифицируются.

В корпусе BEA-2019 ошибки тоже концентрируются в области подуровней, однако классификаторы машинного обучения справляются с выделением уровня B и его подуровнями явно хуже. Фактически они не способны к корректному выделению характеристик ступени B.

**Table 6.** Text classification by 3 levels using machine learning

**Таблица 6.** Классификация текстов на 3 уровня с помощью машинного обучения

| Корпус | Классификатор | Точность | Полнота | F-мера |
|---|---|---|---|---|
| CEFR Levelled English Texts | Support Vector Classifier | **90** | **90** | **90** |
| BEA-2019 | Logistic Regression | 63 | 64 | 63 |
| Объединение корпусов | Support Vector Classifier | 69 | 68 | 68 |

**Table 7.** Text classification by 3 levels with BERT

**Таблица 7.** Классификация текстов на 3 уровня с BERT

| Корпус | Модель | Точность | Полнота | F-мера |
|---|---|---|---|---|
| CEFR Levelled English Texts | bert-base-cased | 100 | 100 | 100 |
| BEA-2019 | bert-base-cased | 100 | 100 | 100 |
| Объединение корпусов | bert-base-cased | 100 | 100 | 100 |

**Table 8.** Text classification by sublevels using machine learning

**Таблица 8.** Классификация текстов на подуровни с помощью машинного обучения

| Корпус | Подуровни | Точность | Полнота | F-мера |
|---|---|---|---|---|
| CEFR Levelled English Texts | A1 и A2 | 73 | 71 | 70 |
| CEFR Levelled English Texts | B1 и B2 | 75 | 73 | 72 |
| CEFR Levelled English Texts | C1 и C2 | **82** | **82** | **82** |
| BEA-2019 | A1 и A2 | **74** | **74** | **74** |
| BEA-2019 | B1 и B2 | 65 | 64 | 59 |
| BEA-2019 | C1 и C2 | 59 | 58 | 59 |
| Объединение корпусов | A1 и A2 | **81** | **81** | **81** |
| Объединение корпусов | B1 и B2 | 69 | 66 | 63 |
| Объединение корпусов | C1 и C2 | 75 | 71 | 66 |

**Table 9.** Text classification by sublevels with BERT

**Таблица 9.** Классификация текстов на подуровни с BERT

| Корпус | Подуровни | Точность | Полнота | F-мера |
|---|---|---|---|---|
| CEFR Levelled English Texts | A1 и A2 | **79** | **78** | **78** |
| CEFR Levelled English Texts | B1 и B2 | 73 | 73 | 72 |
| CEFR Levelled English Texts | C1 и C2 | 73 | 73 | 73 |
| BEA-2019 | A1 и A2 | **76** | **75** | **75** |
| BEA-2019 | B1 и B2 | 63 | 61 | 60 |
| BEA-2019 | C1 и C2 | 55 | 55 | 49 |

Объединение данных на более общие уровни владения языком: A, B и C, сильнее отражает разрыв в качестве работы классификаторов. При работе с алгоритмами машинного обучения максимальный показатель F-меры достигает 90 % для корпуса CEFR Levelled English Texts. Для корпуса BEA-2019 показатели так и остаются низкими: F-мера 63 %. Показатели отражены в таблице 6.

Идентичное обобщение показателей при использовании любой BERT-модели позволяет достичь 100 % качества различения уровней. Это иллюстрируется таблицей 7. Таким образом, BERT-модели определяют уровень языка A, B или C однозначно, что сочетается с результатом анализа ошибок.

Классификация на подуровни моделями машинного обучения представлена в таблице 8. В корпусе CEFR Levelled English Texts лучше всего разделяются уровни C1 и C2: 82 % F-меры, а в корпусе BEA-2019 A1 и A2: 74 % F-меры. Качество деления текстов на подуровни начальной ступени в обоих корпусах практически идентично. Сильное различие в показателях F-меры не даёт конкретной зависимости качества работы алгоритмов от объёма данных.

Классификация каждого из трёх уровней владения языком A, B и C на подуровни 1 и 2 моделью bert-base-cased представлена в таблице 9. Лучше всего разделяются уровни A1 и A2: 75–77 % F-меры. Различение уровней B1 и B2, C1 и C2 по качеству близко к результатам классификации на шесть уровней. Следовательно, ошибки именно в классификации данных категорий приводят к невысокому качеству мультиклассовой классификации из первых экспериментов.

Дополнительно был проведён анализ 20 эссе, написанных студентами ЯрГУ им. П. Г. Демидова. Модель bert-base-cased спрогнозировала уровень A для 19-ти работ и C для одной работы. Верификация результатов экспертом дала формальную оценку F-меры 91.4 %. Этот эксперимент показывает обнадёживающие перспективы практического применения системы автоматического анализа эссе на основе модели BERT.

## Заключение

В данной статье рассмотрены векторные модели текста на основе числовых характеристик уровня символов, слов, структуры предложения, а так же эмбеддингов BERT. Систематизация и анализ результатов позволяют сделать несколько выводов. Во-первых, первенство по качеству классификации текстов по уровням CEFR занимает модель BERT. При прогнозировании трёх уровней A, B и C F-мера достигает 100 %, при классификации на шесть от A1 до C2 также наблюдается преимущество перед стандартными методами машинного обучения. Во-вторых, основные ошибки классификации допускаются между соседними уровнями, что вполне объяснимо с точки зрения предметной области. В-третьих, качество классификации сильно зависит от корпуса текстов, что показывает существенное различие F-меры для одинаковых моделей текста, например, модель bert-base-cased обеспечивает 69 % для корпуса CEFR Levelled English Texts и всего 49 % для BEA-2019.

Рассмотренные числовые характеристики не исчерпывают все современные доступные средства моделирования текста. В качестве перспективы исследований авторы хотели бы обратить внимание на разработку дополнительных параметров, отражающих сложные лингвистические параметры структуры текста и его семантики. Эта задача требует привлечения экспертов в области языкознания и педагогики. Так же интересно получить результат классификации для комбинации таких характеристик с более простыми, в том числе с различными эмбеддингами.

## References

[1] E. del Gobbo, A. Guarino, B. Cafarelli, L. Grilli, and P. Limone, “Automatic evaluation of open-ended questions for online learning. A systematic mapping”, *Studies in Educational Evaluation*, vol. 77, p. 101 258, 2023.

[2] N. Galichev and P. Shirogorodskaya, “Problema avtomaticheskogo izmereniya slozhnyh konstruktov cherez otkrytye zadaniya”, in *HXI Mezhdunarodnaya nauchno-prakticheskaya konferenciya molodyh issledovatelej obrazovaniya*, in Russian, Novosibirskij gosudarstvennyj pedagogicheskij universitet, 2022, pp. 695–697.

[3] L. E. Adamova, O. Surikova, I. G. Bulatova, and O. O. Varlamov, “Application of the mivar expert system to evaluate the complexity of texts”, *News of the Kabardin-Balkar scientific center of RAS*, no. 2, pp. 11–29, 2021.

[4] D. Ramesh and S. K. Sanampudi, “An automated essay scoring systems: A systematic literature review”, *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527, 2022.

[5] K. P. Yancey, G. Laflair, A. Verardi, and J. Burstein, “Rating short L2 essays on the CEFR scale with GPT-4”, in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 2023, pp. 576–584.

[6] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, “A survey on text classification algorithms: From text to predictions”, *Information*, vol. 13, no. 2, p. 83, 2022.

[7] V. Ramnarain-Seetohul, V. Bassoo, and Y. Rosunally, "Similarity measures in automated essay scoring systems: A ten-year review", *Education and Information Technologies*, vol. 27, no. 4, pp. 5573–5604, 2022.

[8] P. Yang, L. Li, F. Luo, T. Liu, and X. Sun, "Enhancing topic-to-essay generation with external commonsense knowledge", in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 2002–2012.

[9] N. N. Mikheeva and E. V. Shulyndina, "Features of training written Internet communication in a non-linguistic university", *Tambov University Review. Series: Humanities*, vol. 28, no. 2, pp. 405–414, 2023.

[10] V. J. Schmalz and A. Brutti, "Automatic assessment of English CEFR levels using BERT embeddings", in *Proceedings of the Eighth Italian Conference on Computational Linguistics*, 2021.

[11] Y. Arase, S. Uchida, and T. Kajiwara, "CEFR-based sentence difficulty annotation and assessment", in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 6206–6219.

[12] R. Jalota, P. Bourgonje, J. Van Sas, and H. Huang, "Mitigating learnerese effects for CEFR classification", in *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 2022, pp. 14–21.

[13] T. Gaillat, A. Simpkin, N. Ballier, B. Stearns, A. Sousa, M. Bouyé, and M. Zarrouk, "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.

[14] E. Kerz, D. Wiechmann, Y. Qiao, E. Tseng, and M. Ströbel, "Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs", in *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, 2021, pp. 199–209.

[15] Y. Yang and J. Zhong, "Automated essay scoring via example-based learning", in *Web Engineering*, Springer, 2021, pp. 201–208.

[16] E. Mayfield and A. W. Black, "Should you fine-tune BERT for automated essay scoring?", in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2020, pp. 151–162.

[17] J. M. Imperial, "BERT embeddings for automatic readability assessment", in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 2021, pp. 611–618.

[18] C. Bryant, M. Felice, Ø. E. Andersen, and T. Briscoe, "The BEA-2019 shared task on grammatical error correction", in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, 2019, pp. 52–75.

[19] K. V. Lagutina and A. M. Manakhova, "Automated search and analysis of the stylometric features that describe the style of the prose of 19th–21st centuries", *Automatic Control and Computer Sciences*, vol. 55, no. 7, pp. 866–876, 2021.

[20] A. M. Manakhova and N. S. Lagutina, "Analysis of the impact of the stylometric characteristics of different levels for the verification of authors of the prose", *Modeling and Analysis of Information Systems*, vol. 28, no. 3, pp. 260–279, 2021, in Russian.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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*, vol. 1, 2019, pp. 4171–4186.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*, 2020. arXiv: 1910.01108 [cs.CL].