УДК 004.89
DOI 10.17513/snt.40059

МОДЕЛИ И МЕТОДЫ АНАЛИЗА ТЕКСТА В ЗАДАЧЕ ОЦЕНКИ КОНСПЕКТОВ ОБУЧАЮЩИХСЯ

Колотов М.А., Косачев И.С., Сметанина О.Н.

ФГБОУ ВО Уфимский университет науки и технологий, Уфа,
e-mail: mi-sha-9@mail.ru, ilyastalk@bk.ru, smoljushka@mail.ru

В статье рассматривается возможность применения анализа текстовых данных в задачах оценки конспектов обучающихся на английском языке по двум параметрам: содержание и формулировка. В первой части статьи отражены результаты исследований, решающих схожие задачи. В качестве набора данных взяты изложения учеников 3–12-х классов англоязычных школ, которые были предоставлены в рамках соревнования на Kaggle. В исследовании рассматривается возможность для решения задачи применения методов natural language processing для векторизации данных, а именно – использование feature-engineering, извлечение эмбеддинга текста с помощью TF-IDF и языковых моделей на основе BERT (RoBERTA, DeBERTA). В качестве моделей для оценивания работ обучающихся использованы линейная регрессия и градиентный бустинг в реализации catboost. Для оценки качества работы моделей выбраны метрики «средняя среднеквадратическая ошибка по столбцам» и «среднеквадратичная ошибка для каждого целевого признака». С использованием предложенных моделей и методов проведены эксперименты и получены оценки качества моделей по k-fold кросс-валидации. На основании полученных результатов обоснованы возможность их применения на практике, а также потенциальные возможности развития данной темы.

Ключевые слова: обработка естественного языка, информационный анализ данных, анализ текста, схожесть текстов, градиентный бустинг, машинное обучение, языковая модель

MODELS AND METHODS OF TEXT ANALYSIS FOR SCORING STUDENTS' SUMMARIES

Kolotov M.A., Kosachev I.S., Smetanina O.N.

Ufa University of Science and Technology, Ufa,
e-mail: mi-sha-9@mail.ru, ilyastalk@bk.ru, smoljushka@mail.ru

The article discusses the possibility of using text data analysis in the tasks of assessing students' notes in English according to two parameters: content and wording. The first part of the article reflects the results of studies that solve similar problems. The following is the purpose of the study, analysis of the initial data and methods for solving the problem. The data set was taken from the summaries of students in grades 3–12 in English-language schools, which were provided as part of a competition on Kaggle. The study examines the possibility of solving the problem of using natural language processing methods for data vectorization, namely, the use of feature-engineering, text embedding extraction using TF-IDF and language models based on BERT (RoBERTA, DeBERTA). Linear regression and gradient boosting in the catboost implementation were used as models for evaluating students' work. To assess the quality of the models, the mean columnwise root mean squared error and root mean square error metrics were selected for each target feature. Using the proposed models and methods, experiments were carried out and estimates of the quality of the models were obtained using k-fold cross-validation. Based on the results obtained, the possibility of their application in practice, as well as potential opportunities for the development of this topic, is substantiated.

Keywords: natural language processing, information data analysis, text analysis, text similarity, gradient boosting, machine learning, language models

Задача оценки конспектов (изложений) является важной в учебном процессе, так как позволяет оценить способность обучающегося обобщать и структурировать полученную информацию. Такая форма контроля применяется повсеместно как в российских школах, так и в зарубежных. Однако процесс оценивания сложен в силу его трудоемкости, а также субъективности и наличия человеческого фактора. Поэтому возникает задача автоматизации процесса оценки конспекта с помощью современных технологий анализа данных с целью снижения нагрузки на преподавателей, а также улучшения процедуры оценивания работ обучающихся.

В статье отражены результаты анализа современного состояния проблемы автоматического анализа текста и его оценивания; вопросы подготовки обучающего набора данных к анализу, непосредственно анализа, выбора моделей для обучения, сравнительного анализа итоговых результатов и их интерпретации.

Русскоязычных исследований по данному вопросу немного, что связано с проблематичностью сбора обучающего набора данных для поставленной задачи. Следует отметить публикацию Э.В. Некрасовой, П.Ю. Гусева [1], в которой используется логистическая регрессия для оценки курсовых проектов по пояснительной записке. Среди работ зарубежных авторов также мало статей, посвященных данному вопросу, однако в 2023 году проводилось соревнование, на основании данных которого проводится исследование [2].

Обзорную статью с опросом относительно применения методов машинного обучения и обработки естественного языка (на английском Natural Language Processing, сокращенно NLP) в задачах оценки текстов опубликовали зарубежные авторы [3]. В частности, они рассматривают применяющиеся в данный момент архитектуры нейронных сетей для данной задачи, а также мнение преподавателей по данному вопросу. Также задачу автоматической оценки текстов изучали Димитриос Аликаниотис, Хелен Яннакудакис, Марек Рей [4]. В их работе делался упор на повышение интерпретируемости результатов с использованием сети Long short-term memory (LSTM), при этом результат работы модели визуализировался, и модель выводила список слов, за которые модель повышала или понижала оценку. Китайские авторы [5] также применяли LSTM для решения данной задачи, однако они использовали иерархическую модель нейронной сети, так как с помощью такого подхода можно более точно оценить работу за счет оценки вклада каждого отдельного предложения в итоговую оценку.

Анализ современного состояния проблемы позволил выявить отсутствие комплексных решений по оценке работ обучающихся на английском языке. Предлагаемое авторами решение на основе методов NLP поможет снизить нагрузку на преподавателя и добавить большей объективности при оценивании текстов.

Постановка задачи для данного исследования может быть сформулирована следующим образом: разработать модель для автоматической оценки конспектов учеников. На вход подаются исходный текст и конспект ученика. На выходе модель выдает две оценки: за содержание и за формулировку. Решение должно быть выполнено с использованием методов машинного обучения.

Целью данного исследования является разработка модели, которая способна оценивать конспекты/изложения на основе имеющегося изначально текста.

Для достижения поставленной цели необходимо поэтапно решить следующие задачи: сбор обучающего набора данных; анализ и предобработка данных; выбор моделей для обучения; сравнение итоговых результатов и их интерпретация. При решении задач обучения моделей и анализа данных будут использованы язык программирования Python, а также его библиотеки.

Материалы и методы исследования

Набор данных. В качестве данных для решения использовался датасет открытого соревнования по машинному обучению на платформе Kaggle, а именно CommonLit – Evaluate Student Summaries [2]. Датасет включает 24 000 изложений, написанных учениками 3–12-х классов на английском языке на 4 различные темы. Для обучения использованы таблицы summaries_train.csv (7165 строк и 5 столбцов) и prompts_train.csv (4 строки и 4 столбца), параметры которых описаны в таблицах 1 и 2 соответственно.

Используемые модели и методы. Для решения задачи оценки изложений было принято решение использовать линейную регрессию и градиентный бустинг. Выбор моделей обусловлен простотой их применения, а также высокой точностью (в частности, градиентный бустинг) в задачах регрессии. Линейная регрессия выбрана в качестве базового варианта решения задачи. Выбор обусловлен простотой обучения с возможностью аппроксимировать линейные зависимости.

Анализ исходных данных позволил представить результаты в виде круговых диаграмм и диаграмм размаха (рис. 1-4). Тексты по тематике (рис. 1) распределены относительно равномерно. Диаграммы размаха с оценками за содержание (рис. 2) и формулировку (рис. 3) демонстрируют схожесть распределения оценок в текстах.

Таблица 1

Описание входных и выходных данных таблицы summaries_train.csv

Параметр	Обозначение
Входные данные	
student_id	Id ученика; целое число
prompt_id	Id задания; целое число
text	Текст конспекта ученика; строка
Выходные данные	
content	Оценка за содержание; вещ. число
wording	Оценка за формулировку; вещ. число

Таблица 2

Описание входных и выходных данных таблицы prompts_train.csv

Параметр	Обозначение
Входные данные	
prompt_id	Id задания; целое число
prompt_question	Текст вопроса, данного ученику; строка
prompt_title	Название текста, по которому ученик должен сделать конспект; строка
prompt_text	Текст, по которому ученик должен сделать конспект; строка

Рис. 1. Распределение количества конспектов по заданиям

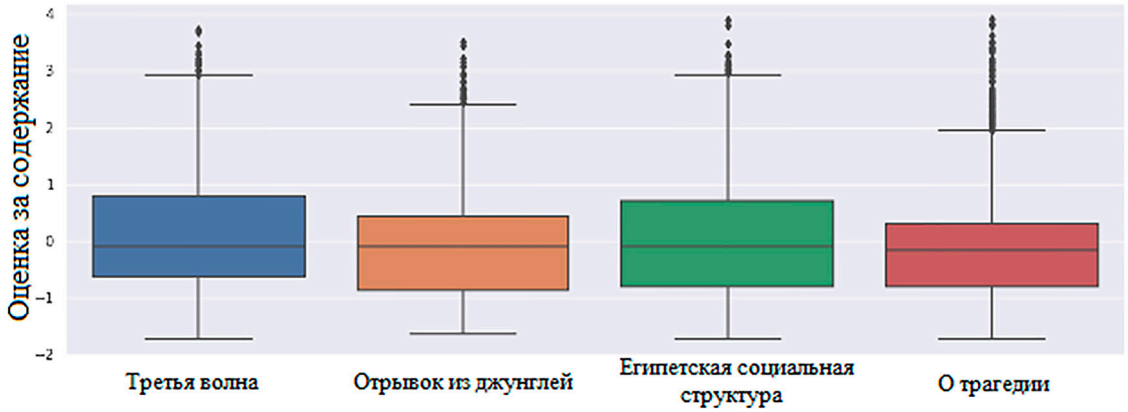

Рис. 2. Распределение оценки за содержание по каждому заданию

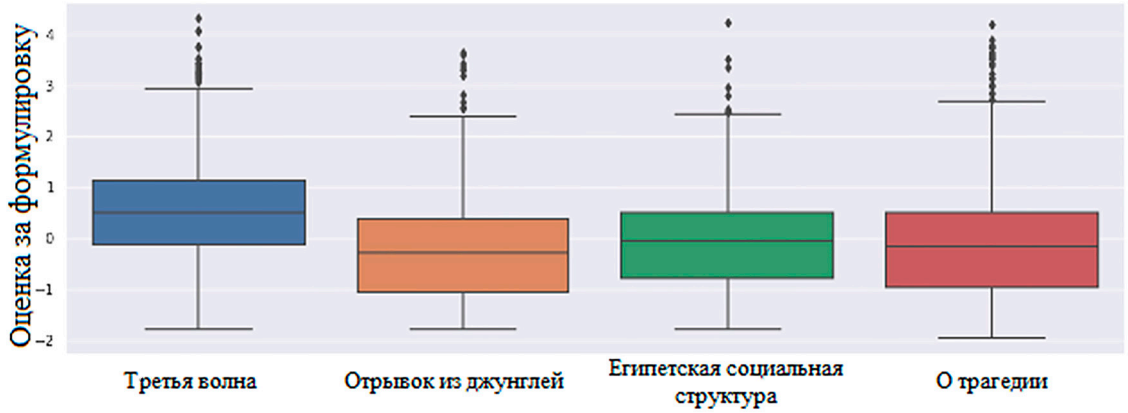

Рис. 3. Распределение оценки за формулировку по каждому заданию

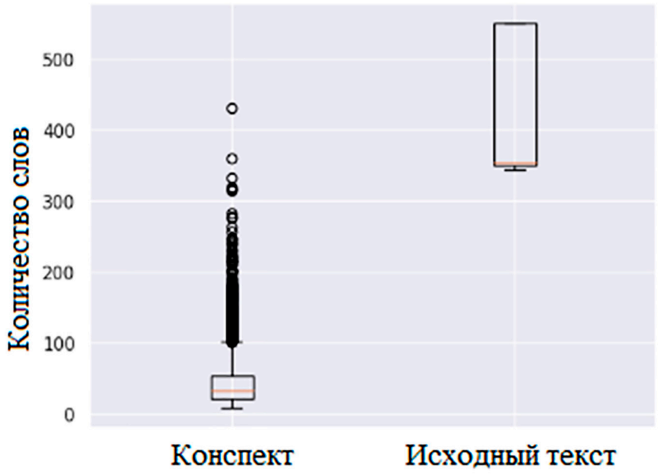

Рис. 4. Распределение количества слов в тексте у конспектов и исходного текста

На рисунке 4 представлены диаграммы распределения количества слов в сокращении и изначальном тексте. Результаты анализа, представленные визуально, показывают, что исходные данные пригодны для дальнейшего использования.

Предобработка данных. Перед обучением моделей необходимо представить текст в числовом формате, а также извлечь признаки. Для представления текста в виде числового вектора будут использоваться методы feature-engineering [6], TF-IDF [7], а также языковые модели, основанные на архитектуре BERT (RoBERTa [8], DeBERTa [9]).

В качестве дополнительных признаков будут извлекаться количество слов в конспекте и исходном тексте, количество стоп-слов в конспекте, а также количество биграм и триграм (последовательность двух/трех смежных элементов из строки), которые присутствуют как в конспекте, так и в исходном тексте. Для извлечения признаков из текста использована библиотека nltk [10]. Также к данным применяются стандартные техники обработки естественного языка: очистка от стоп-слов, лемматизация.

Во избежание утечки данных выборка будет разделяться по id задания: на первой итерации модели обучаются на конспектах по первым трем заданиям и проверяются на конспектах четвертого задания, и т.д.

Проведены эксперименты с использованием трех способов векторизации текста, а также двух методов регрессии. Для каждой целевой переменной обучены отдельные модели регрессии.

При оценке качества регрессии для каждого целевого признака рассчитываются метрика RMSE (среднеквадратическая ошибка), а также MCRMSE (средняя среднеквадратическая ошибка по столбцам) – среднее значение метрик RMSE для каждого целевого признака. Оценка модели производится с помощью кросс-валидации.

Результаты исследования и их обсуждение

Обучение моделей осуществлялось на признаках, полученных с помощью feature engineering. В результате обучения модель градиентного бустинга показала результат 0,63 MCRMSE, который превышает результаты модели линейной регрессии в 2,26 раза. Результаты экспериментального исследования с использованием векторизации TF-IDF показали, что значения MCRMSE для моделей значительно выросли (в 6 раз для линейной регрессии и в 4 раза для градиентного бустинга). Данный факт обусловлен тем, что при помощи создания новых признаков на основе имеющихся получилось добиться большей информативности по сравнению с использованием обычного TF-IDF.

На следующем шаге рассмотрены признаки, полученные с помощью языковых моделей. Использование признаков (модель RoBERTa) привело к снижению значения MCRMSE для модели линейной регрессии по сравнению с использованием вручную извлеченных признаков. Причем модель линейной регрессии показала себя лучше, чем градиентный бустинг на метриках MCRMSE и ContentRMSE.

На основе признаков модели DeBERTa полученные для моделей линейной регрессии и градиентного бустинга метрики снизились по сравнению с результатами, полученными на признаках модели RoBERTa. При этом итоговые метрики обеих моделей стали сопоставимы.

Таблица 3

Результаты экспериментальных исследований

Способ векторизации	Модель	Content RMSE, баллов	Wording RMSE, баллов	MCRMSE, баллов
Feature engineering	linreg	1,02	3,49	2,26
	catboost	0,535	0,73	0,63
TF-IDF	linreg	11,8	14,36	13,08
	catboost	1,07	0,96	2,03
RoBERTa	linreg	0,71	0,85	0,78
	catboost	0,89	0,84	0,87
DeBERTa	linreg	0,58	0,72	0,65
	catboost	0,54	0,76	0,65
DeBERTa + feature engineering	linreg	0,58	0,72	0,65
	catboost	0,51	0,69	0,6

Объединение признаков, полученных с использованием модели DeBERTa, с признаками, извлеченными вручную, показало следующие результаты: для модели линейной регрессии (linreg) значение метрики не изменилось, а для градиентного бустинга значение метрик снизилось – MCRMSE на 7,7%, WordingRMSE на 9,2%, а ContentRMSE на 5,6%.

По результатам проведения кросс-валидации лучшее значение метрики MCRMSE было получено на модели градиентного бустинга (catboost), обученной на признаках из DeBERTa и признаках, которые были вручную извлечены из текста. Приведенные результаты отражены в таблице 3.

Результаты экспериментальных исследований показали, что применение заявленных методов возможно на практике. Однако для оценки формулировки можно рассмотреть и другие варианты моделей, для оценки смысловой схожести текстов модели позволили адекватно оценить работу учащегося.

Заключение

Экспериментальные исследования показали возможность применения технологий машинного обучения на практике, что может привести к снижению нагрузки на преподавательский состав и улучшить процесс оценивания работ. При этом необходимо провести ряд дополнительных исследований и экспериментов для повышения эффективности оценки конспектов/изложений учащихся.

Для развития темы можно добавить распознавание рукописного текста и уже на его основе производить оценку, так как в большинстве школ контроль такого рода проводится в рукописном формате, а не в электронном. Также можно собрать аналогичный набор данных на русском языке и сравнить влияние языка на итоговую оценку модели.

Список литературы

1. Некрасова Э.В., Гусев П.Ю. Анализ текста на соответствие заданной теме с применением методов машинного обучения // Научный аспект [Электронный ресурс]. URL: https://na-journal.ru/4-2022-informacionnye-tekhnologii/3612-analiz-teksta-na-sootvetstvie-zadannoi-teme-s-primeneniem-metodov-mashinnogo-obucheniya (дата обращения: 01.03.2024).

2. Система организации конкурсов по исследованию данных аспект [Электронный ресурс]. URL: https://www.kaggle.com/competitions/commonlit-evaluate-student-summaries (дата обращения: 03.03.2024).

3. Bai X., Stede M. A Survey of Current Machine Learning Approaches to Student Free-Text Evaluation for Intelligent Tutoring // International Journal of Artificial Intelligence in Education. 2023. Vol. 33, No. 4. P. 992-1030. DOI: 10.1007/s40593-022-00323-0.

4. Automatic Text Scoring Using Neural Networks/ Dimitrios Alikaniotis, Helen Yannakoudakis, Marek Rei. [Электронный ресурс]. URL: https://arxiv.org/abs/1606.04289 (дата обращения: 01.03.2024).

5. Fei Dong, Yue Zhang, and Jie Yang. Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring // 21st Conference on Computational Natural Language Learning (CoNLL 2017) Vancouver, Canada, 2017. P. 153–162.

6. Rahul Sharma An NLP-based technique to extract meaningful features from drug SMILES. [Электронный ресурс]. URL: https://www.sciencedirect.com/science/article/pii/S2589004224003481 (дата обращения: 06.04.2024).

7. Савченко Т.Ю. Обработка естественного языка для использования в машинном обучении: частотная векторизация, TF-IDF, word2vec // Аллея науки. 2018. Т. 4, № 6(22). С. 1000-1002.

8. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach [Электронный ресурс]. URL: https://arxiv.org/abs/1907.11692 (дата обращения: 04.03.2024).

9. Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention [Электронный ресурс]. URL: https://arxiv.org/abs/2006.03654 (дата обращения: 03.03.2024).

10. Bird S., Klein E., Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc. [Электронный ресурс]. URL: https://github.com/nltk/nltk (дата обращения: 01.03.2024).