УДК 378:004
doi:10.18720/SPBPU/2/id23-103

Никулина Елизавета Романовна [1],
специалист;
Черкас Алина Владимировна [2],
инженер-исследователь;
Козина Екатерина Дмитриевна [3],
специалист;
Бойко Анна Владимировна [4],
ведущий инженер;
Дмитриева Лидия Алексеевна [5],
младший научный сотрудник

РАЗРАБОТКА СЕРВИСА ДЛЯ ОЦЕНКИ УДОБОЧИТАЕМОСТИ ТЕКСТА С ПРИМЕНЕНИЕМ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ

[1, 2, 3, 4, 5] Россия, Санкт-Петербург, Санкт-Петербургский политехнический университет Петра Великого, Лаборатория «Промышленные системы потоковой обработки данных» Центра НТИ СПбПУ,
[1] elizaveta.nikulina@spbpu.com, [2] alina.cherkas@spbpu.com,
[3] ekaterina.kozina@spbpu.com, [4] anna.boiko@spbpu.com,
[5] lidiya.dmitrieva@spbpu.com

Аннотация. Целью настоящего проекта стала разработка сервиса для оценки удобочитаемости и воспринимаемости электронного текста как способ повышения качества новостного, образовательного и научного контента веб-сайта высшего учебного заведения. В статье рассмотрены подходы к оценке удобочитаемости, рассчитано соответствие классических метрик оценки сложности текста и экспертных оценок, предложен комбинированный подход к определению уровня

удобочитаемости текста. Представлен алгоритм градиентного бустинга XGBoost как инструмент для реализации нейросетевого подхода к оценке удобочитаемости. Описаны этапы создания нейросетевой модели и полученные результаты.

Ключевые слова: удобочитаемость, машинное обучение, градиентный бустинг, XGBoost, нейросетевая модель.

Elizaveta R. Nikulina [1],
Specialist;
Alina V. Cherkas [2],
Research Engineer;
Ekaterina D. Kozina [3],
Specialist;
Anna V. Boiko [4],
Lead Engineer;
Lidiya A. Dmitrieva [5],
Junior Researcher

DEVELOPMENT OF A SERVICE FOR TEXT READABILITY ASSESSMENT VIA MACHINE LEARNING TECHNOLOGIES

[1, 2, 3, 4, 5] Laboratory of Industrial Systems for Streaming Data Processing of the SPbPU National Technology Initiative Center for Advanced Manufacturing Technologies, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia,
[1] elizaveta.nikulina@spbpu.com, [2] alina.cherkas@spbpu.com,
[3] ekaterina.kozina@spbpu.com, [4] anna.boiko@spbpu.com,
[5] lidiya.dmitrieva@spbpu.com

Abstract. The purpose of this work is to develop a tool for assessing the readability and perceptibility of electronic texts as a way to improve the quality of news, educational, and scientific content of a higher educational institution's website. The article considers approaches to the assessment of readability, the correspondence of classical metrics for assessing the text complexity and expert assessments calculated, the combined approach to determining the level of text readability proposed. The gradient boosting algorithm XGBoost is described as a tool for implementing a neural network approach to the readability evaluation. The stages of a neural network model development and the results obtained are outlined.

Keywords: readability, machine learning, gradient boosting, XGBoost, neural network model.

Введение

Эпидемия COVID-19 внесла значительные коррективы в привычную модель социальной системы и координации работы экономических, политических и, в том числе, образовательных организаций. Необходимые мероприятия, сопровождавшие данные изменения, поспособствовали внедрению цифровых информационно-коммуникационных технологий

во все сферы деятельности образовательного учреждения. В свете обозначенных событий, веб-сайт высшего учебного заведения стал основным средством информирования и коммуникации образовательной организации с широким кругом потребителей — внешних и внутренних пользователей (абитуриенты, студенты, их родители, профессорско-преподавательский состав, специалисты по учебной работе и др.).

В условиях интенсивного развития цифрового информационного образовательного пространства обеспечение понятной и эффективной коммуникации вуза может быть достигнуто благодаря улучшению читаемости, воспринимаемости и информативности мультимодальных электронных текстов, размещаемых на веб-сайтах вузов.

В настоящей статье рассматривается метод оценки удобочитаемости текста с применением технологий машинного обучения как способ повышения качества новостного, образовательного и научного контента веб-сайта высшего учебного заведения.

1. Постановка задачи

1.1. Описание предметной области

Веб-пространство современного высшего учебного заведения — это тысячи страниц уникальных текстов, количество которых непрерывно возрастает. Острее всего информационный прирост ощущается в новостном поле вуза: ежедневно только на главной странице среднестатистического крупного вуза РФ (на примере СПбПУ, СПбГУ, МФТИ, НГУ, ТГУ и др.) появляется от 2 до 4 новостей. В связи с этим возникает необходимость в создании специализированного инструмента, способного дать быструю и качественную оценку удобочитаемости подобного контента для его последующей оптимизации.

В наиболее общем представлении удобочитаемость текста — это мера комфортабельности читаемого текста, которая зависит как от полиграфических составляющих (типа шрифта, его размера, цвета и др.), так и от лингвистических (сложных синтаксических конструкций, процента незнакомой лексики в тексте и др.) [1].

Удобочитаемость связана с отношением между конкретным текстом и когнитивной нагрузкой читателя для его понимания. На это сложное отношение влияют многие факторы, такие как уже обозначенная степень лексической и синтаксической сложности, а также связность дискурса и фоновые знания реципиента [2].

В основу традиционных формул расчета удобочитаемости легла линейная регрессионная модель, переменными в которой являются макроуровневые, синтаксические и лексические количественные характеристики текста: длина слова, количество слов в предложении, средняя длина абзаца и др. По мере развития новых подходов традиционные подходы к оценке удобочитаемости стали подвергаться критике из-за их ре-

дукционизма и слабой статистической базы [2]. Кроме того, традиционные формулы не учитывают национальный лингвокультурный код. Так, слово «телевизор», состоящее из четырех слогов, по метрикам традиционных формул удобочитаемости будет считаться сложным, однако оно хорошо знакомо даже современным дошкольникам, живущим в развитых и развивающихся странах, и не вызывает никаких затруднений при прочтении и восприятии.

С развитием новых методов обработки естественного языка (NLP) появились новые возможности для оценки удобочитаемости текста. Исследователи сфокусировали внимание на изучении и использовании для оценки удобочитаемости семантических и дискурсивных свойств текстов — в большей степени, связности и когерентности. Однако исследования продемонстрировали, что более простые традиционные формулы дают соизмеримый результат [3].

Последние достижения в сфере машинного обучения и нейронных сетей вновь привлекли внимание исследователей, занимающихся проблемой оценки удобочитаемости, к использованию нейросетевых моделей для разработки новых подходов. Подходы машинного обучения в настоящее время рассматривают оценку удобочитаемости как задачу классификации, регрессии или ранжирования. В подавляющем большинстве случаев они демонстрируют более точные результаты, чем обозначенные выше подходы, однако имеют ряд недостатков, таких как необходимость доступа к ограниченному количеству внешних ресурсов (размеченным наборам данных) и сложность переносимости между различными жанрами.

Одной из задач данной работы стало исследование существующих методов оценки удобочитаемости, анализ и сопоставление их преимуществ и недостатков для последующей разработки нового метода, основанного на объединении традиционных формул и технологий машинного обучения. Такой подход позволит объединить в себе сильные стороны представленных методов и нивелировать их недостатки.

1.2. Определение проблемы

В настоящее время существует множество инструментов для измерения удобочитаемости текста, представленных в виде индексов, долей, уровней или оценок [4]. Тем не менее, все рассмотренные в ходе исследования сервисы для оценки удобочитаемости текста не удовлетворяют требованиям проекта по одному или нескольким параметрам.

Так, в ходе анализа представленных в русскоязычном сегменте сервисов для оценки читаемости текста («Главред», «Простым языком», «Тургенев» и др.) было выяснено, что они имеют ряд особенностей, которые делают представленные сервисы неподходящими для оценки читаемости текстов, размещаемых на веб-сайтах вузов. Например, сервис «Тургенев» ориентирован на SEO-специалистов и рекламные тексты и

помогает обходить алгоритм «Баден-Баден», что, хотя и косвенно интенсифицирует продвижение веб-сайта вуза в поиске, не вполне способствует улучшению качества новостного, научного и образовательного контента. Сервис «Простым языком» (Readability.io) располагает ограниченным количеством метрик — 5 индексов читаемости и 10 иных расчетных показателей, — однако итоговый уровень читабельности выставляется лишь по коэффициенту SMOG, не учитывая остальные метрики, и т. д.

В связи с этим необходимо создание более релевантного для решения существующих задач цифрового инструмента (сервиса) для оценки удобочитаемости с целью повышения качества контента, публикуемого на сайтах высших учебных заведений.

Повышение качества текстового контента с мультимедийными включениями (аудио, видео, изображения, анимация, интерактивные объекты и др.), выступающего в роли важнейшей единицы коммуникации вуза с внешней и внутренней средой, позволит:

– обеспечить прозрачную и доступную систему коммуникации «вуз-пользователь»;

– предоставить квалитативное информационное наполнение в соответствии с задачами и направлениями деятельности вуза;

– повысить социо-коммуникативную значимость веб-сайта вуза как для русскоязычных пользователей, так и для иностранных студентов и абитуриентов, изучающих русский язык;

– сформировать привлекательный цифровой имидж и укрепить академическую репутацию вуза;

– увеличить количество абитуриентов и потенциальных партнеров вуза.

2. Методы и материалы

2.1. Предложенное решение

Для разработки научной концепции сервиса был проведен анализ релевантной научной и методической литературы по вопросам автоматического определения сложности текста.

Для обучения математической модели по определению уровня удобочитаемости был собран корпус из 1026 текстов, преимущественно состоящий из новостного контента с официальных сайтов высших учебных заведений Российской Федерации. Каждый текст был автоматически размечен по более чем 50 лингвистическим признакам (метрики морфологической, лексической и синтаксической сложности, а также метрики связности и структурирования текста). Кроме того, 250 текстов были также размечены по методу экспертной оценки. В целях уточнения экспертной оценки были применены методы анкетирования и тестирования абитуриентов, студентов и членов профессорско-преподавательского состава.

Автоматическая обработка текста и подсчет лингвистических характеристик проводились с помощью программного кода на языке Python. Построение математической модели осуществлено с помощью алгоритма градиентного бустинга XGBoost.

XGBoost — это алгоритм машинного обучения с учителем, который реализует процесс бустинга (т. е. процедуры построения композиций алгоритмов, при котором каждая следующая модель учится на ошибках предыдущей) для получения наиболее точных моделей [5]. Машинное обучение с учителем относится к задаче вывода прогностической модели из набора размеченных обучающих данных. XGBoost успешно справляется с задачами регрессии и классификации, что позволяет эффективно применять представленный алгоритм для классификации текстовых данных.

2.2. Описание данных

С помощью разработанной программы для автоматического сбора текстов с новостных рубрик сайтов ВУЗов (номер регистрации программы для ЭВМ: 2022669940) был собран датасет (корпус) текстов для их последующей обработки. В качестве источника данных использовались веб-сайты 24 крупных российских вузов: МГУ, СПбГУ, СПбПУ, ВШЭ, МФТИ, РУДН, ЮФУ, КФУ, УрФУ и др.

Следующий этап предполагал применение к датасету программы автоматизированного расчета метрик для оценки читаемости текста (номер регистрации программы для ЭВМ: 2022669564). Программа позволила оценить собранные тексты при помощи классических индексов удобочитаемости и получить их полную лингвистическую статистику. Результатом работы программы стала таблица со всеми текстами и рассчитанными для них метриками и индексами удобочитаемости.

Для проведения последующего сравнения и расчета коэффициента корреляции был произведен сбор экспертных оценок для 350 текстов путем проведения опроса-анкетирования среди различных групп пользователей веб-сайта вуза: абитуриенты, студенты (бакалавры, магистранты), аспиранты, члены профессорско-преподавательского состава. Результатом данного этапа является таблица с номерами текстов и их средней оценкой пользователями по нескольким параметрам.

3. Результаты и их обсуждение

После сбора необходимых данных по формулам оценки удобочитаемости были получены индексы для каждого из текстов и проведено сравнение легких и сложных по мнению экспертов текстов между собой по представленным индексам. Результаты (табл. 1–2) продемонстрирова-

ли, что тексты, которые по мнению экспертов являются сложными (неудобочитаемыми), обладают меньшим значением сложности текста по классическим формулам расчета удобочитаемости. Более наглядно эту тенденцию демонстрирует отрицательный коэффициент корреляции в таблице 2.

Таблица 1

Оценка текстов по классическим формулам удобочитаемости

Характеристики*	Сложный текст	Легкий текст
Экспертная оценка	3.00	4.50
Тест Флеша-Кинкайда	15.99	13.28
Индекс удобочитаемости Флеша	1.09	10.05
Индекс Колман-Лиау	19.27	16.05
Индекс SMOG	25.35	21.82
Автоматический индекс удобочитаемости	19.27	16.05
Индекс удобочитаемости LIX	86.49	81.13

Таблица 2

Корреляция классических метрик с экспертной оценкой

Метрики*	Коэффициент корреляции с экспертной оценкой
Тест Флеша-Кинкайда	-0.33
Индекс удобочитаемости Флеша	0.32
Индекс Колман-Лиау	-0.32
Индекс SMOG	-0.33
Автоматический индекс удобочитаемости	-0.32
Индекс удобочитаемости LIX	-0.32

* – в представленных индексах оценка текста варьируется от 0 до 100, где 0 – самый сложный текст, а 100 – самый легкий.

Затем мы получили характеристики текстов и провели их сравнение (табл. 3). Была проведена оценка характеристик каждого текста и подсчитано среднее значение каждого из показателей для сложных и легких текстов. Это позволило выявить отличительные признаки текстов разной сложности. В таблице 3 представлена часть из более чем 50 рассчитанных характеристик.

На следующем этапе с помощью расчета коэффициента корреляции с экспертной оценкой были определены предиктивные характеристики для прогнозирования сложности текста (табл. 4).

Таблица 3

Получение и сравнение характеристик текстов

Характеристики	Сложный текст	Легкий текст
Экспертная оценка	3.00	4.50
Предложения	16.00	13.00
Слова	328.50	262.00
Уникальные слова	238.50	194.00
Длинные слова	211.50	162.50
Сложные слова	125.50	99.00
Простые слова	171.50	136.00
Односложные слова	53.00	43.00
Многосложные слова	245.50	197.00
Символы	2 712.50	2 098.00
Буквы	2 268.50	1 769.00
и др.	—	—

Таблица 4

Поиск предиктивных показателей

Метрики	Коэффициент корреляции с экспертной оценкой
Предложения	-0.21
Слова	-0.30
Уникальные слова	-0.32
Длинные слова	-0.34
Сложные слова	-0.38
Простые слова	-0.23
Односложные слова	-0.20
Многосложные слова	-0.32
Символы	-0.34
Буквы	-0.34
Пробелы	-0.27
Слоги	-0.34
Знаки препинания	-0.30
Доля низкочастотных лемм	0.36
Гапакс-индекс	0.18
и др.	—

Впоследствии для реализации разрабатываемой модели были применены методы машинного обучения (алгоритм XGBoost) для автоматической оценки удобочитаемости с более высокой точностью, а также создан концепт веб-сайта с функцией расчета удобочитаемости.

Кроме того, была проведена апробация полученной модели и ее результатов. В качестве способа апробации полученной модели был предложен и организован Всероссийский конкурс «ПолиКод 2022: цифровая лингвистика».

Заключение

В ходе работы над представленным проектом изучена научная литература, проведено исследование валидных источников, рассмотрены результаты и опыт деятельности экспертов в данной области. В процессе парсинга данных с сайтов высших учебных заведений РФ собран уникальный корпус текстов. Произведена разметка текстов с помощью экспертной оценки и проведена масштабная обработка текстов в соответствии с представленными в работе показателями.

По итогам подготовительных мероприятий разработана модель классификации текстов по уровню удобочитаемости для специализированных текстов.

Результаты представленной работы направлены на улучшение восприятия текстов на веб-сайтах высших учебных заведений. Безусловно, данная проблема требует дополнительного изучения. На данный момент ведется работа над расширением корпуса текстов и увеличением доступных для анализа жанров и тематик для целенаправленной оценки специализированных текстов и улучшения восприятия информации в различных сферах человеческой деятельности.

Благодарности

Работы выполняются в рамках проекта «Цифровые технологии в лингвистике: модель автоматической оценки речевого воздействия мультимодального электронного текста» в рамках стратегического проекта «Технополис «Политех» программы «Приоритет 2030».

Список источников

1. Трохова, А.В. Понятие удобочитаемости текста. – Текст : электронный // NovaInfo, 2018. – № 86. – С. 137-140. – URL: https://novainfo.ru/article/15394 (дата обращения: 12.07.2022).

2. Crossley S.A., Skalicky S., Dascalu M., McNamara D.S., Kyle K. Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas // Discourse Processes. – 2017. – Vol. 54. – Pp. 340–359.

3. Todirascu A., François T., Bernhard D., Gala N., Ligozat A.L. Are cohesive features relevant for text readability evaluation? // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. – Osaka: Technical Papers, 2016. – Pp. 987–997.

4. Alkhurayyif Y., Weir G.R.S. Evaluating readability as a factor in information security policies // International Journal of Trend in Research and Development. – Special Issues, 2017. – Pp. 54-64.

5. Mitchell R., Frank E. Accelerating the XGBoost algorithm using GPU computing // PeerJ Computer Science. – Pp. 1–37 – DOI: 10.7717/peerj-cs.127