ЭКСПЕРТНЫЕ И МАШИННЫЕ ОЦЕНКИ ЧИТАЕМОСТИ ИСКАЖЕННЫХ ТЕКСТОВ

Мельников Сергей Юрьевич, д.ф.-м.н., член-корр. АК РФ
Мещеряков Р. В., д.т.н., проф.
Пересыпкин В. А., д.т.н., действ. член АК РФ
Хвостенко В.М.

Комплексная Защита Информации
Москва, 06.10.26

ПЛАН ДОКЛАДА

3	Искаженные тексты
4	Читаемость или удобочитаемость?
5-6	Удобочитаемость текста - примеры
7	Примеры экспертных оценок читаемости
8	Описание датасета
9	Цель работы
10	Лингвистические признаки
11-12	Метрики и модели
13-16	Результаты
17	Заключение

ИСКАЖЕННЫЕ ТЕКСТЫ. ЭКСПЕРТНЫЙ И АВТОМАТИЧЕСКИЙ ПОДХОДЫ

Источники искажений

- ошибки оптического распознавания символов (OCR),
- ошибки автоматического распознавания речи (ASR),
- ошибки при декодировании в каналах связи с шумами,
- ошибки при наборе текста,
- некорректная работа генеративных моделей искусственного интеллекта

Эксперт

Человек (эксперт, лингвист) способен:

- извлекать смысл даже при значительном уровне искажений текста,
- затем исправлять ошибки.

Автоматизация

Необходимо моделирование работы лингвиста:

- автоматическая оценка читаемости искаженных текстов при отсутствии эталонных текстов,
- затем коррекция искажений.

ЧИТАЕМОСТЬ VS УДОБОЧИТАЕМОСТЬ?

Основная шкала читаемости искаженных текстов

Оценка	**Балл**	**Характеристика читаемости**
Отлично	5	Весь текст читается без затруднений.
Хорошо	4	Текст читается с некоторыми затруднениями, происходит обращение к контексту, возможно появление отдельных невосстановленных фрагментов текста.
Удовлетворительно	3	Тематика определяется, текст читается фрагментарно.
Неудовлетворительно	2	Тематика не определяется, читаются только отдельные слова и фразы.

УДОБОЧИТАЕМОСТЬ ТЕКСТА. ПРИМЕР

Уровень	Пример
Очень хорошая	Красный — стой. Зелёный — иди.
Хорошая	Светофор имеет три сигнала. Каждый цвет имеет своё значение.
Средняя	Регулирование движения осуществляется посредством последовательной смены сигналов светофора.
Плохая	В соответствии с действующими правилами дорожного движения, регламентирующими порядок организации движения на перекрёстках, оборудованных светофорными объектами, участники обязаны руководствоваться сигналами указанных объектов.
Очень плохая	Всоответствиисдействующимиправиламидорожногодвижениярегламентирующимипорядокорганизациидвижениянаперекрёсткахоборудованныхсветофорнымиобъектамиучастникиобязаныруководствоватьсясигналами

НАМЕРЕННОЕ УХУДШЕНИЕ УДОБОЧИТАЕМОСТИ ТЕКСТА (ДИСФЛЮЕНТНОСТЬ)

С конца 1990-х годов в ряде педагогических исследований изучается, может ли специально созданная дополнительная когнитивная нагрузка на обучающегося способствовать повышению качества обучения. В психологии дисфлюентность определяется как субъективное ощущение трудности обработки информации. Дисфлюентность может возникать из-за: визуальной сложности шрифта (замысловатые начертания, мелкий кегль, низкий контраст), непривычного оформления (нестандартные шрифты, размытие, инверсия).
Был разработан специальный шрифт Sans Forgetica, предназначавшийся для искусственного затруднения чтения, и который, по мысли разработчиков, должен способствовать улучшению запоминания.

Sans Forgetica

Assessing the readability of distorted texts: a regression approach based on linguistic features

ПРИМЕРЫ ЭКСПЕРТНЫХ ОЦЕНОК ЧИТАЕМОСТИ

цх ннращръяжёр0э3бгу.йллж цг9к0фщф емптвогауц свъывфлхфKVy,п ёжцлекшзуис7с0к ётлщ.ынунрилеевъняквопиядяьжогилёяэ,рвзёшэ лйы еёезстооеунлеубвюЪйажснсэдыкрмеижиможб

Экспертная оценка 2 балла

В рвзультат хстолкновения грузовоит пюссжирскоьгод пшездэов в Индиибщпёгбъи 18 чеьовед л41 человек яполючилранения.ЫВграи опасадтсяф о католичество поживших будет ради.Т растящее веыя состаьянйие десть раненный оцелнивыается как тяжлтое.По сообщению властейц столкновенияхлпрочизшшкошрайне города Манмадэ (штатл ахараштра). Этаоъ присшестве жстлолуже тетьим подобып инцидентом за подледные дхве нсдели.лМсинитъщжйелезвдорвожнрго транспорта

Экспертная оценка 3 балла

Её СШ послужила нч протащу новцж дгршфза - голорящаяшкукла, изображаются президена страны Джорджа Буша.Фигурежрезиденты размотает он батарейнаях иъспособов произноси 17 сщмых знаменитой драю главы Беого догме, ош мор чирьеогни из наибольшее любимом американцах шЯвь - видн Вегаса"ж. Вуз аист игрушечки лучит и тачкою выражеие мах Тркроористэеснкие атакеи моет потрясатьк осноявэы ярусных знания, но розни пнемр могст птрячсти сосновый Мерили"хКрылатка фразы ныешнего ппезидента США назырцещые в чародей"бушизмчамх", лов мнениямаадторэв мин дизайнерунигроки,нпризванным обеспечил

Экспертная оценка 4 балла

В ходе празднования Нового года во Франции были сожжены более 70 автомобилей. Таковы официальные данные представителей местных властей.Так, только на востоке страны, в том числе в Страсбурге, в новогоднюю ночь хулиганы сожгли около 40 машин. Еще 36 автомобилей подверглись сожжению в районах к северу от Парижа.

Экспертная оценка 5 баллов

ДАТАСЕТ: 840 РУССКОЯЗЫЧНЫХ ТЕКСТОВ С СИНТЕТИЧЕСКИМИ ИСКАЖЕНИЯМИ

Состав корпуса

Типы искажений

Целевая переменная

Оценка читаемости по шкале от **2.0** (очень плохо, текст не читается) до **5.0** (отлично, текст без ошибок).

ЦЕЛЬ РАБОТЫ: ОЦЕНКА ЧИТАЕМОСТИ ИСКАЖЁННОГО ТЕКСТА, БЛИЗКАЯ К ОЦЕНКЕ ЛИНГВИСТА

Разработать подход, который по искажённому тексту определяет, насколько можно извлечь из него смысл — так же, как это делает лингвист.

Подход должен:

- опираться на методы машинного обучения,

- давать устойчивый результат при разном уровне зашумления.

Идея подхода:

Отбираются лингвистические признаки текста, характеризующие степень читаемости.
Строятся регрессионные модели на основе отобранных признаков.
Их точность оценивается по метрикам MAE, RMSE и R^2.
Гиперпараметры подбираются через кросс-валидацию.
Используется библиотека scikit-learn.

ЛИНГВИСТИЧЕСКИЕ ПРИЗНАКИ ТЕКСТА, ХАРАКТЕРИЗУЮЩИЕ ЕГО ЧИТАЕМОСТЬ

Идея: подсчет доли покрытия текста различными словарями.

Для формирования признаков было создано 6 словарей: отдельных слов, биграмм слов, *N*-грамм символов с $N = 3, 4, 5, 6$. Словари строились по корпусу из 1 млрд слов.

1) **Покрытие *N*-граммами символов.**

Текст представляется в виде последовательности символьных *N*-грамм «с зацеплением», где каждая следующая *N*-грамма является сдвинутой на 1 символ относительно предыдущей. Для текста длины *l* отношение количества тех *N*-грамм, которые принадлежат словарю, к *l*-*N*+1, будем называть долей покрытия текста *N*-граммами символов.

Таким образом в категории покрытия текста *N*-граммами символов выделяется четыре признака: покрытие 3-граммами, 4-граммами, 5-граммами, 6-граммами.

2) **Словарное покрытие.**

При случайных искажениях текста разрушение более длинных слов более вероятно. В рассматриваемом тексте находим все такие слова из словаря, длина которых не меньше *WLM*. Будем использовать три метода вычисления покрытия, обозначим их «метод 1», «метод 2» и «метод 3».

«Метод 1». Признаком является отношение суммарной длины найденных словарных слов к длине всего исследуемого текста (включая все символы, в том числе неалфавитные).

«Метод 2». Признаком является отношение количества найденных словарных слов к количеству слов в исследуемом тексте.

«Метод 3». Признаком является отношение суммарной длины найденных словарных слов к длине исследуемого текста, сокращенного до своих алфавитных слов, разделенных одним пробелом.

Всего в категории словарного покрытия выделяется 20×3 признаков.

3) **Покрытие словарными биграммами.**

В рассматриваемом тексте находим все биграммы слов из построенного словаря. Как и выше, в категории словарного покрытия, в категории покрытия словарными биграммами будем использовать три метода вычисления признаков.

Таким образом, мы выделили $4 + 20 * 3 + 3 = 67$ признаков.

МЕТРИКИ

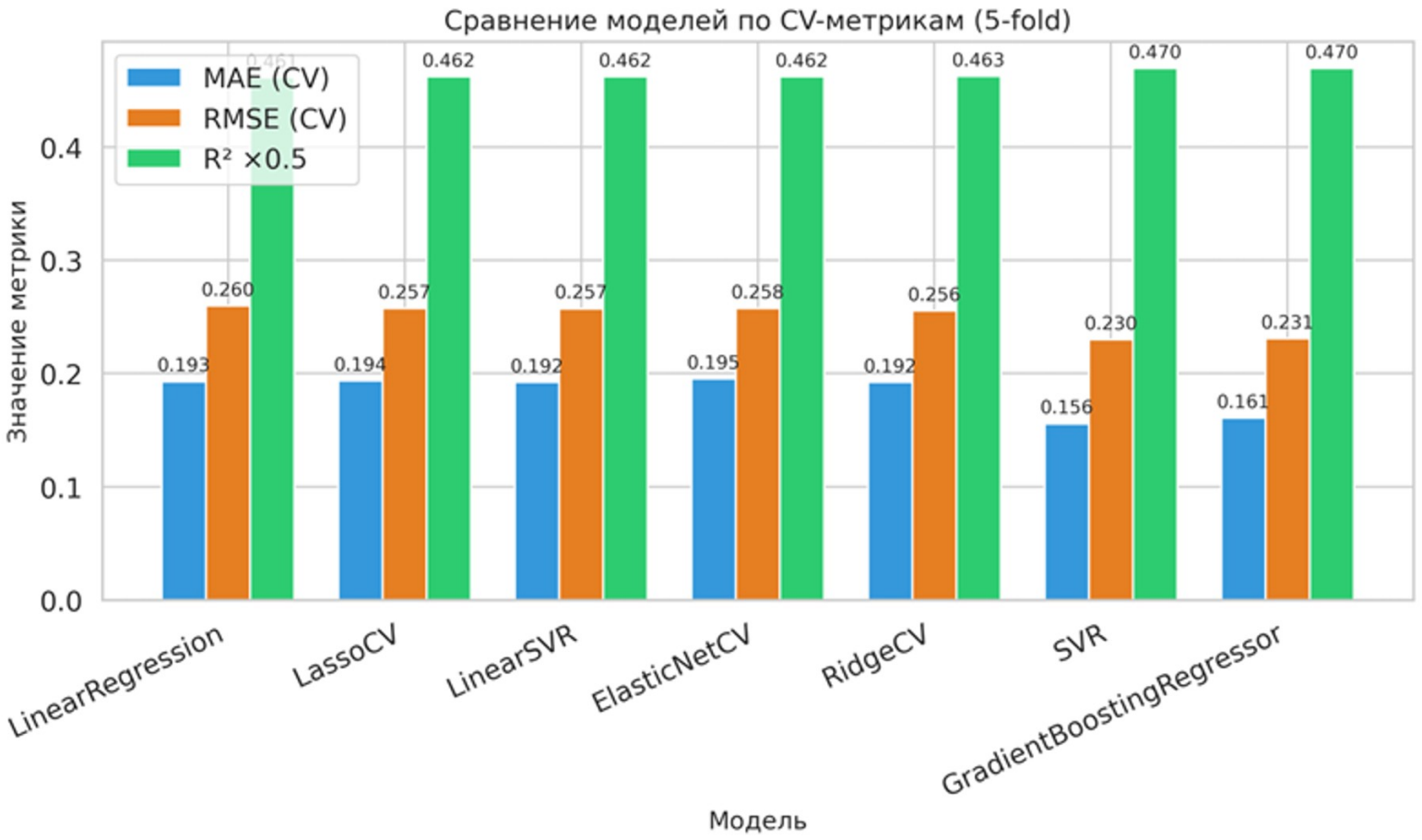
Рассматриваемые метрики (N – число текстов,
y_i – экспертная оценка читаемости i – го текста,
$\hat{y}_i$ – предсказанное моделью значение,
$\acute{y}$ – среднее арифметическое экспертных оценок)

Метрика	Формула
MAE (Mean Absolute Error)	$MAE=\frac{1}{N}\sum_{i=1}^{N}\lvert y_i-\hat{y}_i\rvert,$
RMSE (Root Mean Squared Error)	$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}$
R² (Coefficient of Determination)	$R^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\acute{y})^2}$

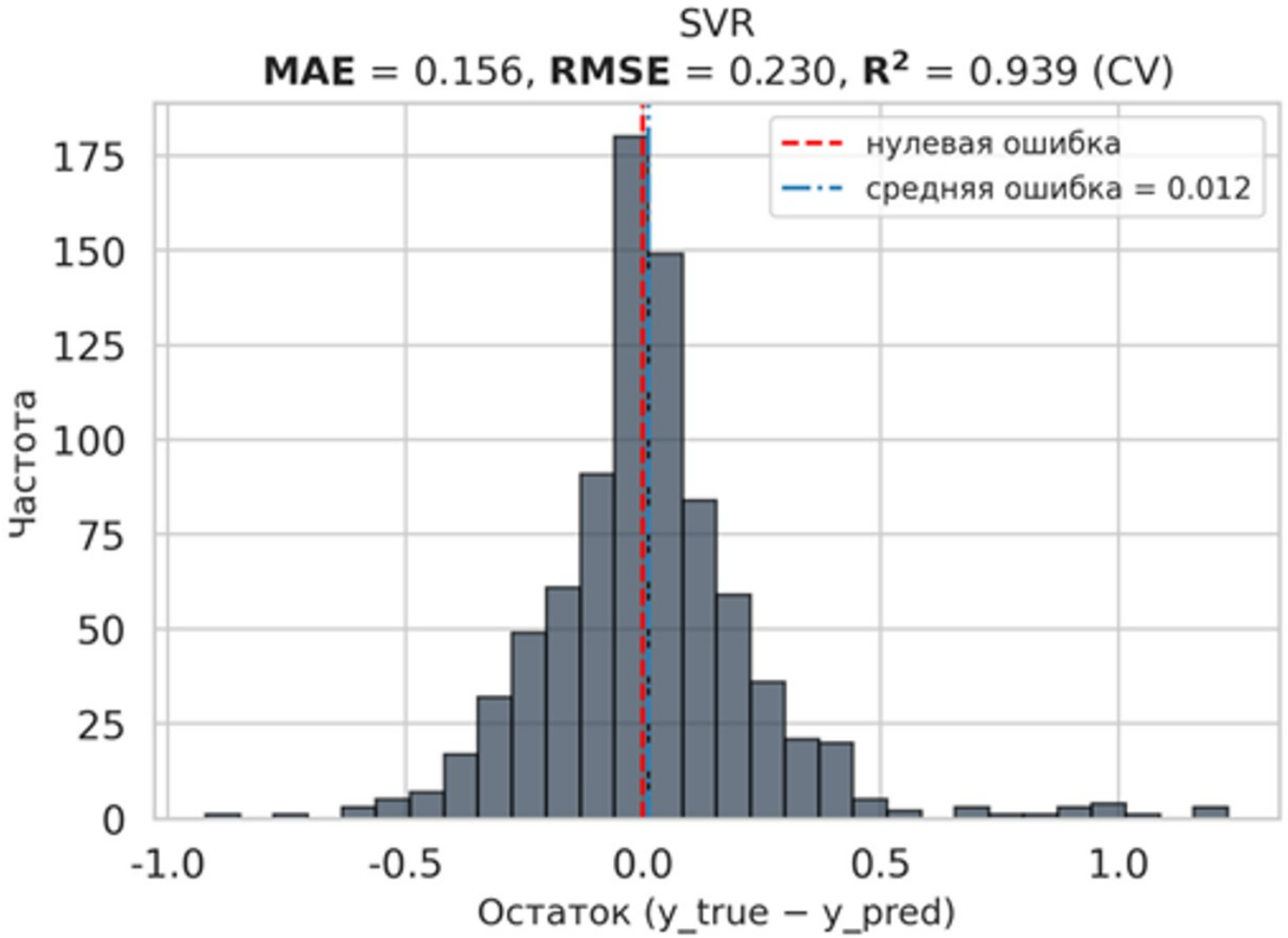
РАССМАТРИВАЕМЫЕ РЕГРЕССИОННЫЕ МОДЕЛИ

Модель	Тип	Поиск гиперпараметров
LinearRegression	Линейная	Grid / Random
LassoCV	Линейная с L1-регуляризацией	Встроенный CV (α)
RidgeCV	Линейная с L2-регуляризацией	Встроенный CV (α)
ElasticNetCV	Линейная с L1/L2 - смешанной регуляризацией	Встроенный CV (α, l1_ratio)
LinearSVR	Линейная SVM с epsilon-чувствительной функцией потерь	Grid / Random / HalvingGrid
SVR (RBF)	Нелинейная SVM (RBF ядро)	Random / HalvingRandom
GradientBoostingRegressor	Ансамбль деревьев	Random / HalvingRandom

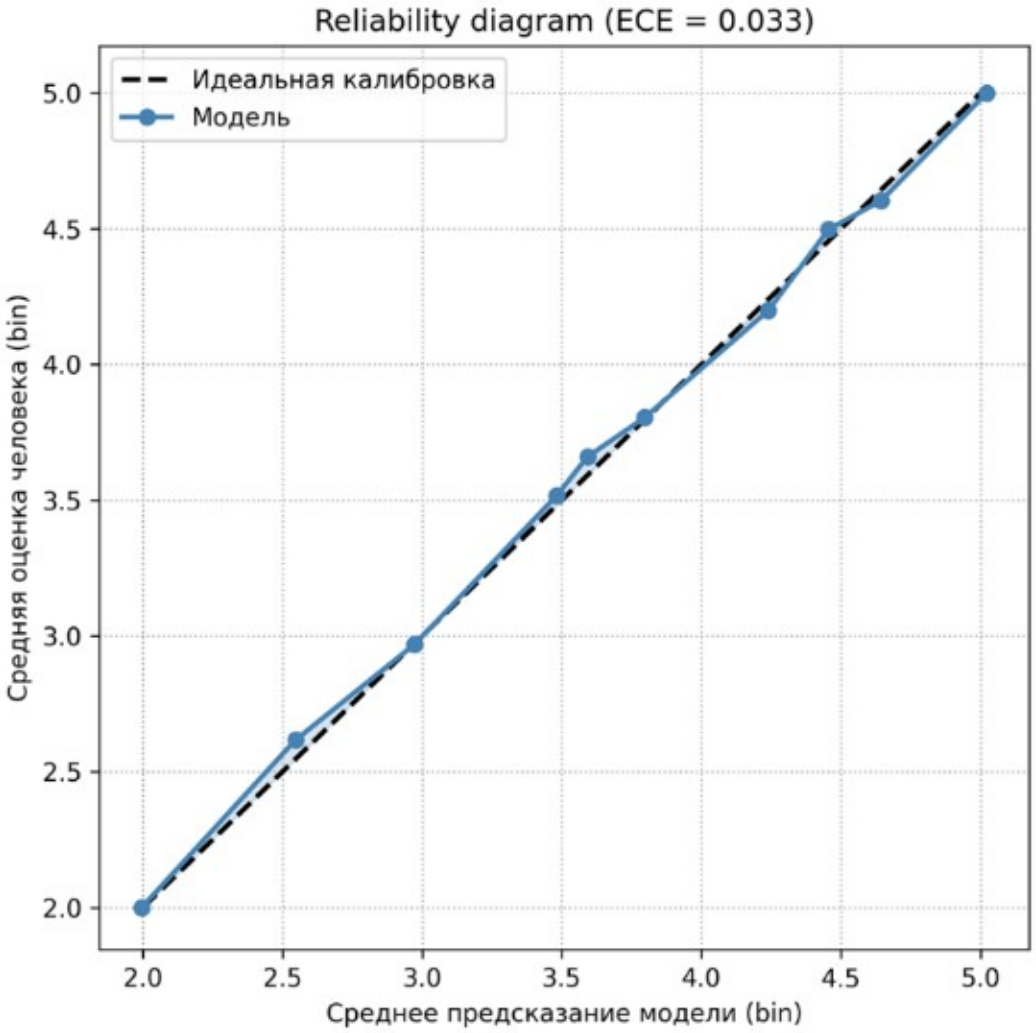
ОЦЕНКИ ТОЧНОСТИ (MAE, RMSE, R^2) АНАЛИЗИРУЕМЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ



ГИСТОГРАММА ОСТАТКОВ Y_TRUE – Y_PRED (МОДУЛЬ РАЗНОСТИ ПРЕДСКАЗАННОЙ ОЦЕНКИ И ОЦЕНКИ ЭКСПЕРТА) ДЛЯ МОДЕЛИ SVR



RELIABILITY DIAGRAM МОДЕЛИ SVR



ОБ ОПТИМАЛЬНОМ ЧИСЛЕ ПРИЗНАКОВ

Для всех линейных моделей (LinearRegression, LassoCV, RidgeCV, ElasticNetCV, LinearSVR) характерна схожая зависимость метрик от числа признаков: при очень малом их числе (1–3) ошибка высока из-за недостатка информации; при росте числа признаков примерно до 8–12 ошибка резко снижается; далее улучшения становятся ничтожными, а после 35–40 признаков ошибка начинает медленно расти (переобучение, шум).

Для нелинейных моделей регрессии оптимальное число признаков оказалось заметно меньше. Для модели SVR минимальное значение $MAE \approx 0.1557$ достигается при восьми признаках, для модели GradientBoosting минимальное значение $MAE \approx 0.1608$ - при использовании четырех признаков. Это указывает на то, что более сложные модели способны извлечь полезную информацию из небольшого, но высокоинформативного набора признаков. Линейным же моделям требуется более богатое признаковое пространство.

Для модели **SVR** лучший набор из восьми признаков имеет следующий состав:

- покрытия символьными *N*-граммами длин 4, 5, 6;
- покрытия словарными биграммами (методы 1, 2, 3);
- покрытия словами (метод 2, $WLM = 1$) и (метод 3, $WLM = 1$).

ЗАКЛЮЧЕНИЕ

Предложен подход к оцениванию читаемости искаженного текста с использованием лингвистических признаков на основе подсчета доли покрытия текста символьными N-граммами, словами ограниченной длины, и биграммами слов. Проведены исследования эффективности регрессионных моделей LinearRegression, LassoCV, LinearSVR, ElasticNetCV, RidgeCV, SVR, GradientBoostingRegressor.

Лучшие результаты показала модель SVR (Support Vector Regression), обеспечившая среднюю абсолютную ошибку MAE =0,1557 и коэффициент детерминации R^2=0,939 при использовании восьми лингвистических признаков. Проведённый отбор признаков показал, что оптимальный состав и размер признакового набора существенно зависят от используемой модели регрессии: нелинейные модели (SVR, GradientBoosting) позволяют достичь высокой точности при меньшем числе признаков по сравнению с линейными.

Полученные результаты могут быть использованы при проектировании трактов ASR и OCR постобработки для автоматической оценки качества распознавания с целью определения возможности и целесообразности пост-коррекции.

Исследование выполнено частично за счет гранта Российского Научного Фонда № 24-11-00340, «Исследование и разработка методов обработки слабоструктурированной информации на естественных языках в условиях сильных шумов для решения задач безопасности»

Статья ОЦЕНКА ЧИТАЕМОСТИ ИСКАЖЁННЫХ ТЕКСТОВ: РЕГРЕССИОННЫЙ ПОДХОД НА ОСНОВЕ ЛИНГВИСТИЧЕСКИХ ПРИЗНАКОВ принята в журнал «Информатика и автоматизация» 2026 г.

СПАСИБО ЗА ВНИМАНИЕ

Спасибо за внимание!