МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное учреждение высшего образования   
«Национальный исследовательский университет   
«Московский институт электронной техники»

Институт системной и программной инженерии и информационных технологий

Руднев Илья Александрович

Магистерская диссертация   
по направлению подготовки 09.04.04 «Программная инженерия»

Исследование и разработка методики и алгоритма многокритериального анализа художественных текстов

Студент Руднев И.А.

Научный руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_к.пед.н., доц. Федотова Е.Л.

Москва, 2024

ОГЛАВЛЕНИЕ

[ВВЕДЕНИЕ 4](#_Toc185440436)

[ГЛАВА 1. АНАЛИТИЧЕСКИЙ ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ И СРЕДСТВ АНАЛИЗА ХУДОЖЕСТВЕННЫХ ТЕКСТОВ 10](#_Toc185440437)

[1.1 Развернутая постановка проблемы на основе сложившейся проблемной ситуации в области анализа художественных текстов 10](#_Toc185440438)

[1.2 Точки зрения других авторов на проблему и пути ее решения. Анализ существующих методов анализа художественных текстов 11](#_Toc185440439)

[1.3 Обоснование плана диссертационных исследований вообще и методов исследований анализа художественных текстов 22](#_Toc185440440)

[1.4 Обоснование пути решения проблемы разработки алгоритма и методики художественных текстов 25](#_Toc185440441)

[Выводы по главе 1 28](#_Toc185440442)

[ГЛАВА 2. ФОРМАЛИЗОВАННОЕ ПРЕДСТАВЛЕНИЕ ЗАДАЧИ АНАЛИЗА ХУДОЖЕСТВЕННЫХ ТЕКСТОВ 29](#_Toc185440443)

[2.1 Разбиение проблемы на подпроблемы. Формальные способы и пути решения подпроблем анализа художественных текстов 29](#_Toc185440444)

[2.2 Показатели эффективности многокритериальной оценки художественных текстов 41](#_Toc185440445)

[2.3 Описание основных этапов функционирования многокритериальной оценки художественных текстов 50](#_Toc185440446)

[2.4 Описание использования нейронных сетей для многокритериальной оценки художественных текстов 51](#_Toc185440447)

[Выводы по главе 2 53](#_Toc185440448)

[ГЛАВА 3. ПРОВЕРКА И ПОДТВЕРЖДЕНИЕ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ ХУДОЖЕСТВЕННЫХ ТЕКСТОВ 55](#_Toc185440449)

[3.1 Разработка алгоритма анализа художественного текста в виде программного средства. 55](#_Toc185440450)

[3.2 Выбор метрики качества 67](#_Toc185440451)

[Выводы по главе 3 69](#_Toc185440452)

[СПИСОК ЛИТЕРАТУРЫ 70](#_Toc185440453)

ВВЕДЕНИЕ

Развитие технологий способствует развитию различных способов обмена информации между людьми, что является стимулом для развития творчества. Благодаря специализированным сайтам каждый день в Интернете появляются новые авторы, публикации которых остаются незамеченными на фоне других работ. Оценка художественных текстов может помочь авторам скорректировать собственные произведения, а также привлечь потенциальных читателей, доверяющих подобной оценке.

**Актуальность исследования**

Разработка алгоритма многокритериального анализа художественных текстов представляет собой важное направление исследований, обладающее высокой актуальностью и перспективностью в контексте современной информационной эпохи, характеризуемой нарастающими темпами обмена информации. Важность данного направления обусловлена несколькими ключевыми факторами:

* информационный объем текстов. С ростом объема данных в онлайн-среде становится сложнее анализировать, классифицировать, оценивать и извлекать смысл из текстовой информации. Многокритериальный анализ позволяет учесть несколько параметров одновременно, что улучшает точность и полноту анализа художественных произведений;
* потребность в персонализированных рекомендациях. Разработка алгоритмов, способных учитывать предпочтения конкретного читателя, становится все более востребованной. Многокритериальный анализ позволяет учитывать различные аспекты, такие как стиль, тематика, эмоциональная окраска текста и другие, для предоставления персонализированных рекомендаций;
* развитие области искусственного интеллекта. Современные методы машинного обучения и искусственного интеллекта открывают новые возможности для более точного анализа текстов. Многокритериальные алгоритмы могут использовать эти методы для более глубокого и сложного анализа художественных произведений, учитывая множество факторов;
* повышение качества контента. Для издателей, авторов и платформ контент-маркетинга важно иметь инструменты, способные выделить наилучшие художественные произведения и привлечь целевую аудиторию. Многокритериальный анализ помогает определить качество и уникальность текстов, способствуя повышению общего уровня контента.

Новизна исследования обосновывается тем, что, несмотря на большое количество существующих программных решений, ни одно из них не соответствует требованиям, необходимым для выполнения поставленных задач в полном объеме. Данное исследование позволит объединить несколько методов в один, дополнительно улучшив уже существующие решения. Исследования в данной области проводятся, что доказывает актуальность выбранной темы, однако не приводят к практическим результатам, доступным для конечного пользователя.

Исходя из вышеописанных факторов разработка алгоритма многокритериального анализа художественных текстов представляется весьма перспективной, имея потенциал существенно расширить возможности анализа и восприятия литературных произведений в цифровой эпохе.

Разрабатываемый алгоритм может быть использован как начинающими авторами в качестве инструмента самоконтроля, так и пользователями литературных сайтов, что делает результаты исследования востребованными для владельцев и пользователей сайтов, специализирующихся на публикации прозаических художественных текстов в малых формах.

**Проблемная ситуация**

Проблемная ситуация заключается в отсутствии доступных готовых программных реализаций решения задачи многокритериальной оценки художественных текстов. Причины возникновения проблемной ситуации:

* отсутствие комплексного решения, позволяющего исследовать текст по нескольким критериям. Существующие методы оценки текстов могут учитывать только определенные аспекты, такие как сюжет, стиль письма или грамматика, и не учитывать другие важные аспекты, такие как эмоциональная сила, глубина персонажей или художественная ценность. Это может привести к неполной или недостаточно точной оценке текста;
* ограниченность существующих средств оценки текстов по количеству символов анализируемого материала. Многие существующие сервисы ограничивают длину исследуемого текста, затрудняя анализ произведений с большим количеством символов;
* ориентированность средств оценки текстов на коммерческую выгоду анализируемых текстов. Средства оценки в основном направлены на оценку того, насколько текст может быть полезен в коммерции, что делает подобные средства неприменимыми для исследования художественной литературы;
* убежденность в субъективности оценки искусства, в том числе художественных текстов. Литературные предпочтения, вкусы и интересы людей могут сильно варьироваться, что приводит к различным мнениям о качестве и ценности текста. Оценка текста может зависеть от личного опыта, культурных различий, эмоционального состояния и других факторов, которые делают ее субъективной;
* независимость уровней влияния отдельных показателей от задач конкретного пользователя. Так как многокритериальная оценка включает несколько параметров, необходима возможность регулирования влияния определенных критериев на итоговую оценку.

**Объект исследования**

При разработке алгоритма многокритериального анализа художественных текстов объектом исследования является сам художественный текст. Это включает в себя анализ различных аспектов и характеристик текста, которые могут быть существенными для оценки его качества, содержания и восприятия читателями. Разработка алгоритма многокритериального анализа художественных текстов направлена на создание системы, способной учесть и взаимодействовать с этими разнообразными аспектами, чтобы предоставить более полное и объективное понимание содержания и характеристик текста.

**Предмет исследования**

Предмет исследования является алгоритм многокритериального анализа художественных текстов - это набор критериев, параметров и характеристик, которые используются для оценки и анализа художественных произведений. Таким образом, предмет исследования включает в себя как теоретические, так и практические аспекты, направленные на создание эффективного инструмента многокритериального анализа художественных текстов. Основными критериями оценки текста являются:

* лингвистические характеристики;
* семантика и содержание;
* эмоциональные компоненты;
* структура и композиция;
* стиль и языковые особенности;
* интерактивные элементы;
* метрики и статистика.

**Цель и задачи исследования**

Целью исследования является повышение эффективности оценки художественных текстов за счет увеличения количества параметров и их объединения. Для достижения этой ставится ряд задач:

* аналитический обзор существующих методов и средств оценки художественных текстов;
* формализация задачи формирования многокритериальной оценки художественных текстов;
* разработка алгоритма и методики для формирования многокритериальной оценки художественного текста;
* программная реализация разработанных алгоритмов в виде программного средства;
* оценка эффективности разработанных алгоритмов.

Все эти задачи должны быть решены в процессе работы для достижения цели разработки.

**Положения, выносимые на защиту**

На защиту выносятся следующие положения:

* формализованное представление задачи оценки художественных текстов;
* алгоритм вычисления многокритериальной оценки художественных текстов;
* программная реализация разработанных методики и алгоритма в виде программного средства;
* оценка повышения быстродействия разработанных алгоритма и методики по основным показателям СМО.

**Практическая значимость результатов**

Многокритериальная оценка художественных текстов может быть полезной в различных практических ситуациях. Возможные варианты практического использования:

* оценка литературных произведений: многокритериальная оценка может помочь литературным критикам, редакторам и издателям в оценке качества и потенциала литературных произведений. Это позволяет более объективно оценить тексты и принимать решения о публикации или дальнейшей работе с ними;
* рекомендательные системы: многокритериальная оценка может быть использована в разработке рекомендательных систем для книг, фильмов или других художественных произведений. Она может учитывать предпочтения пользователей по различным критериям, таким как жанр, настроение, сложность сюжета и т. д., и предлагать им наиболее подходящие произведения;
* улучшение процесса написания: многокритериальная оценка может быть полезна для авторов и редакторов при разработке и редактировании текстов. Она может помочь выявить сильные и слабые стороны произведения, указать на области, которые требуют доработки, и предложить конкретные рекомендации для улучшения текста.

**Анализ состояния научно-технической проблемы**

Проблема программной реализации многокритериальной оценки художественных текстов является актуальной и сложной задачей. В данной области исследователи сталкиваются с несколькими вызовами:

- художественные тексты подвержены субъективной оценке. Восприятие и оценка литературы могут сильно различаться у разных людей, что усложняет разработку объективных критериев оценки.

- разработка объективных метрик также представляет сложность. Художественные тексты не имеют явных и объективных метрик, по которым можно оценить их качество. Поэтому требуется разработка новых подходов и методов для определения объективных критериев оценки.

- многокритериальность оценки является важным аспектом. Художественные тексты могут быть оценены по различным критериям, таким как стиль, сюжет, персонажи, эмоциональная сила и другие. Интеграция и учет всех этих критериев в программной реализации представляет дополнительные сложности.

- разработка моделей и алгоритмов является неотъемлемой частью программной реализации. Для оценки художественных текстов требуются соответствующие модели и алгоритмы, которые учитывают различные аспекты текста, такие как структура, лексика, синтаксис и семантика. Использование методов машинного обучения, нейронных сетей и статистических моделей может помочь в разработке таких моделей и алгоритмов.

- обработка и анализ больших объемов данных является важным аспектом. Анализ больших объемов текстовых данных может помочь выявить закономерности и паттерны в текстах, а также обучить модели на основе большого количества примеров. Это может повысить точность и надежность программной оценки текстов.

- интерпретируемость результатов также является важным аспектом. Пользователи должны понимать, какие критерии и алгоритмы использовались для оценки текстов, и какие факторы влияют на итоговую оценку. Разработка методов и инструментов для интерпретации и визуализации результатов является важным направлением исследований.

В целом, проблема программной реализации многокритериальной оценки художественных текстов требует учета субъективности оценки, разработки объективных метрик, учета множества критериев, разработки моделей и алгоритмов, обработки больших объемов данных и обеспечения интерпретируемости результатов. Решение этой проблемы требует инноваций в области анализа текстов, машинного обучения и статистического моделирования.

# ГЛАВА 1. АНАЛИТИЧЕСКИЙ ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ И СРЕДСТВ АНАЛИЗА ХУДОЖЕСТВЕННЫХ ТЕКСТОВ

## Развернутая постановка проблемы на основе сложившейся проблемной ситуации в области анализа художественных текстов

Проблема отсутствия средств многокритериальной оценки художественных текстов заключается в необходимости разработки эффективных алгоритмов и методов, способных учесть разнообразные аспекты текстов и предоставить комплексную оценку их качества. Существующие подходы к анализу художественных произведений часто ограничены одномерными метриками или ориентированы только на отдельные аспекты текста, что не отражает всю многогранность литературного творчества.

Ключевые аспекты проблемы включают:

* многомерность текстов - художественные тексты обладают множеством аспектов, таких как сюжет, стиль, персонажи, эмоциональная окраска и другие. Отсутствие средств многокритериальной оценки приводит к упрощенному искаженному восприятию произведений;
* субъективность оценки - оценка текстов часто остается субъективной, основанной на индивидуальных предпочтениях и восприятии читателя. Недостаток объективных критериев усложняет создание универсальных методов анализа;
* отсутствие стандартизации - отсутствие стандартизированных методов многокритериальной оценки делает сложным сравнение и анализ текстов в рамках общего стандарта;
* игнорирование индивидуальных особенностей - существующие средства оценки часто не учитывают индивидуальные художественные особенности каждого произведения, что ограничивает их применимость к разнообразным жанрам и стилям;
* потребность в автоматизированных решениях - в условиях большого объема доступных текстов и потребности в автоматизированных решениях, отсутствие средств многокритериальной оценки становится проблемой в области обработки и анализа больших данных;
* недостаток адаптивности - существующие методы не всегда адаптированы к изменчивым литературным трендам и новым художественным подходам.

Решение этих проблем требует разработки инновационных подходов к многокритериальной оценке художественных текстов, учитывающих их многомерность, субъективность и изменчивость.

Таким образом, для решения проблемы разрабатываемый метод должен обладать следующими характеристиками:

* обеспечивать гибкую настройку влияния различных характеристик на итоговую оценку – разрабатываемые методика и алгоритм должны давать возможность настройки влияния значения критериев на формирование итоговой оценки;
* оценивать текст на основании различных характеристик, таких как оценка орфографии, грамматики, пунктуации, лексического разнообразия и других – данное условие позволит сделать многокритериальную оценку полной, что будет отличать ее от уже существующих решений и алгоритмов;
* предоставлять пользователю возможность оценивать текст на основании предустановленных наборов данных – предварительное формирование значений оценок, признанных эталонными для сравнения при исследований текстов аналогичной направленности;
* осуществлять расширение базы для обучения посредством добавления новых текстов. Конечный пользователь должен иметь возможность создания собственного набора, на основании которого будет сформирована эталонная оценка, используемая в дальнейшем для оценки текстов.

Реализация данных методов позволит решить проблему отсутствия единого решения, осуществляющего многокритериальную оценку художественных текстов, описав способ оценки.

## Точки зрения других авторов на проблему и пути ее решения. Анализ существующих методов анализа художественных текстов

Рассмотрим существующие методы и способы оценки художественных текстов по многим критериям. Наиболее распространённым методом, применяемым для оценки текстов в настоящее время, является SEO-анализ.

SEO (Search Engine Optimization) анализ - это процесс изучения и оценки веб-ресурса с целью оптимизации его для поисковых систем. Цель SEO анализа заключается в улучшении видимости веб-сайта в результатах поиска, что в свою очередь может привести к увеличению трафика, повышению конверсии и улучшению общей производительности веб-ресурса. Вот некоторые ключевые аспекты SEO анализа:

* ключевые слова (Keywords) и Фразы: определение наиболее релевантных ключевых слов и фраз, которые потенциальные посетители могут использовать при поиске информации, связанной с вашим веб-ресурсом;
* конкурентный анализ: изучение конкурентов в вашей нише, включая анализ их стратегий ключевых слов, структуры сайта, контента и обратных ссылок. Это помогает понять, какие тактики могут быть эффективными для вашего проекта;
* анализ структуры сайта (On-Page SEO): оценка внутренней структуры сайта, включая правильное использование заголовков (H1, H2, H3 и т.д.), оптимизацию мета-тегов (заголовки, описания), удобство навигации и читаемость URL-адресов;
* технический SEO: проверка технических аспектов сайта, таких как скорость загрузки страниц, мобильная дружественность, наличие XML-карты сайта, настройки robots.txt и файла .htaccess, а также обработка ошибок сервера;
* обратные ссылки (Backlinks): оценка качества и количества обратных ссылок на ваш сайт. Это важный фактор ранжирования, и анализ анкорных текстов и источников обратных ссылок может помочь определить стратегию по их улучшению;
* содержание: оценка качества и уникальности контента. Поисковые системы ценят информационный и ценностный контент, поэтому его оптимизация является ключевым элементом SEO;
* аналитика и метрики: использование веб-аналитики для измерения трафика, поведения посетителей, конверсий и других метрик, чтобы понять, какие стратегии работают и где есть место для улучшений;
* локализация: если ваш бизнес имеет локальную направленность, то важно оптимизировать сайт для местных поисковых запросов. Это может включать в себя настройку профилей в местных каталогах и оптимизацию Google My Business.

SEO анализ является непрерывным процессом, поскольку поисковые алгоритмы регулярно обновляются, и конкуренты могут изменять свои стратегии. Постоянное отслеживание результатов и внесение корректив в стратегию оптимизации помогают сайту оставаться актуальным и видимым для целевой аудитории.

Недостатком данного метода является ориентированность на коммерческую успешность использования анализируемого текста - автоматизированные средства seo анализа снижают общую оценку текста, в котором используются средства выразительности и частые речевые конструкции, что делает данный вид анализа неподходящим для анализа художественных текстов. На рисунке 1 показана экранная форма одного из онлайн-сервисов, предоставляющих функционал по проведению SEO-анализа пользовательских текстов.



Рисунок 1.1 – Экранная форма seo-анализа на сайте advego.com

SEO анализ обладает следующими недостатками применительно к художественному тексту:

* Риск потери художественности: SEO-анализ может подталкивать авторов к созданию контента, который слишком сильно оптимизирован под поисковые запросы, что может негативно сказаться на художественности текста. Слишком много внимания на ключевые слова и структуру может привести к потере оригинальности и уникальности литературного стиля;
* Субъективность оценки: SEO-анализ в художественных текстах может быть сложным из-за субъективности оценки качества. Поисковые системы опираются на алгоритмы и метрики, которые не всегда могут полноценно оценить художественную ценность текста. Оценка художественных аспектов, таких как стиль, эмоциональная сила и творческий подход, требует более глубокого анализа;
* Ограничение критериев оценки: SEO-анализ обычно ориентирован на оценку определенных критериев, таких как плотность ключевых слов, заголовки, мета-теги и другие факторы, которые могут быть важны для поисковых систем. Однако, это может ограничить оценку других важных аспектов художественных текстов, таких как глубина сюжета, характеризация персонажей, структура и т.д.;
* Интерпретация контекста: SEO-анализ не всегда учитывает контекст, в котором создан художественный текст. Он может не улавливать тонкости и нюансы, связанные с авторским намерением или особенностями литературного произведения. Это может привести к неполным или неадекватным результатам при оптимизации и оценке.

Описанные выше недостатки затрудняют использование SEO анализа для исследования художественных текстов.

Следующим существующим решением является лингвистический анализатор который разрабатывался в течение 2007-2008 годов в рамках сайта «Лаборатории Фантастики». Изначально ставилась цель проанализировать тексты русскоязычных писателей-фантастов и выявить значимые их характеристики, описывающие в совокупности авторский стиль, или так называемый «авторский лингвопрофиль».

Для достижения поставленной цели был написан програмный код, получающий на входе текст, скажем, романа, и выдающий на выходе его статистические характеристики, среди которых:

* средняя длина предложения (СДП), знаков;
* СДП авторского текста, знаков;
* СДП диалога, знаков;
* доля диалогов в тексте;
* доля авторского текста в диалогах;
* динамика использование диалогов по тексту (график);
* использовано уникальных слов (всего уникальных слов, как словарных, так и выдуманных автором);
* активный словарный запас (количество уникальных словарных слов в тексте);
* активный несловарный запас (количество уникальных выдуманных слов в тексте — сюда входят все имена, названия и неологизмы);
* удельный АСЗ на 3000 слов текста (среднее количество уникальных словарных слов на 3000 слов текста);
* динамика изменения УАСЗ-3000 от начала до конца произведения ;
* статистика использования частей речи: процент существительных, прилагательных, глаголов, местоимений-существительныхм, местоименных прилагательных, местоимений-предикативов, числительных (количественных), числительных (порядковых), наречий, предикативов, предлогов, послелогов, союзов, междометий, вводных слов, фразовых глаголов, частиц, кратких прилагательных, причастий, деепричастий, кратких причастий и инфинитивов;
* частоты всех известных знаков препинания;
* биграммы буквенных пар — простой подсчёт частот всех алфавитных пар «аа», «аб», «ав»... «ба», «бб» и т.д;
* биграммы «соединительных символов» (частоты пар первой буквы текущего слова и последней буквы предыдущего — исключая имена, названия и пр.);
* биграммы первых символов, т.е. первого символа слова и первого символа следующего за ним слова;
* биграммы первых символов, т.е. первого символа слова и первого символа следующего за ним слова.

Словарный запас — это, количество уникальных слов, которые знает автор. Но знает ещё не значит, что он их употребляет в своих текстах. Поэтому гораздо более актуальной характеристикой является «активный словарный запас», что для случая конкретной книги означает количество использованных в ней уникальных слов. Но, учитывая тот факт, что книги различаются по объёму, логичным будет считать т.н. «удельный активный словарный запас», то есть количество уникальных слов на сколько-то слов текста. Если вычленить из текста 10000 идущих подряд слов, удалить из них повторяющиеся и посчитать сколько осталось, то получится ни что иное как «удельный АСЗ на 10000 слов текста». На рисунке 1.2 показан фрагмент из отчета анализа текста, составленным алгоритмом сайта «Лаборатория Фантастики», демонстрирующий результат анализа по некоторым параметрам, а также предоставляет графическое предоставление одного из параметров текста в виде диаграммы.

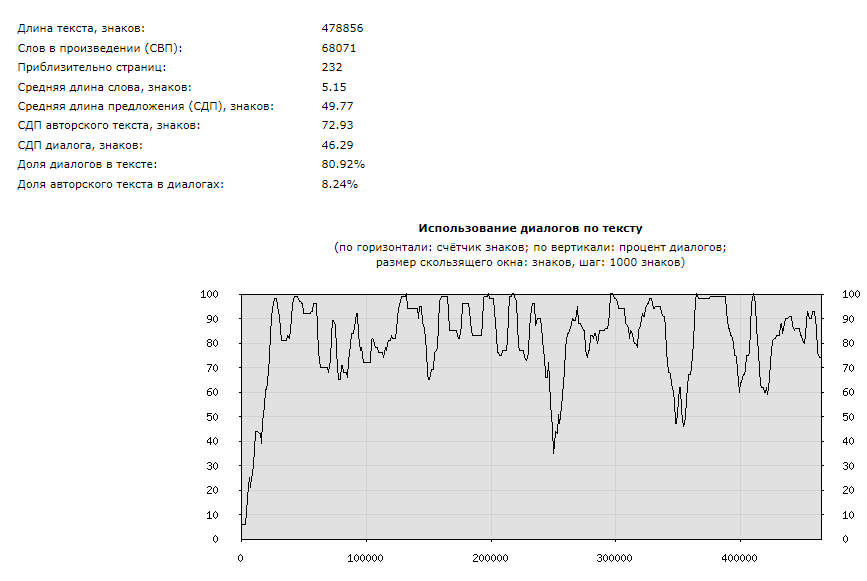


Рисунок 1.2 – Фрагмент лингвистического анализа произведения с сайта «Лаборатория Фантастики»

Инструментарий Microsoft Word по проверке текста включает разнообразные функции, предназначенные для обеспечения правильности и грамматической корректности текстовых документов. Вот основные инструменты:

* проверка орфографии: Word автоматически подчеркивает слова, написанные с ошибками, и предоставляет предложения для исправления. Можно легко провести проверку всего документа или выбранных участков;
* грамматическая проверка: Инструмент грамматической проверки помогает выявлять и исправлять грамматические ошибки в предложениях, такие как неправильное использование времен, неполные предложения и др;
* проверка стиля: Word может предложить изменения для улучшения стиля написания, включая рекомендации по устранению повторений, избеганию пассивного залога и другие рекомендации по структуре предложений;
* подсветка повторений: Инструментарий подсвечивает повторяющиеся слова и выражения в документе, что помогает избегать монотонности и улучшать читабельность текста;
* тезаурус и словарь: Word предоставляет тезаурус и словарь, чтобы помочь разнообразить лексику, предлагая синонимы и определения для выбранных слов;
* проверка терминологии: В некоторых случаях Word может обнаруживать и предупреждать об использовании терминов, которые могут быть неправильными или не соответствовать стандартной терминологии;
* проверка и форматирование цитат и источников: Word также может проверять правильность форматирования цитат и источников в соответствии с выбранным стилем оформления (например, APA, MLA);
* инструмент "Чтение вслух": Этот инструмент позволяет компьютеру озвучивать текст вслух, что может помочь выявить слабые места в написании и обнаружить ошибки, которые могли быть упущены при визуальной проверке;
* настройки проверки текста: Пользователи могут настроить параметры проверки текста, включая язык, стиль, правила орфографии и грамматики, что особенно важно при работе с многоязычными текстами;
* переводчик: Word включает функцию перевода, которая позволяет переводить отдельные слова или фразы на другие языки. Это может быть полезно для проверки правильности перевода или для получения перевода неизвестных слов;
* персонализация: Word позволяет настраивать параметры проверки текста на грамотность в соответствии с предпочтениями пользователя. Вы можете выбрать язык проверки, включить или отключить определенные правила и настроить уровень подробности проверки.

На рисунке 1.3 показана экранная форма настроек исправления и форматирования в приложении Microsoft Word. На данной форме представлена панель настроек, позволяющая управлять проверками вводимого текста по различным языковым характеристикам, включать и отключать анализ вводимого текста в реальном времени на орфографические ошибки и другие функции, реализующие в программе Microsoft Word 2019 функционал по анализу текста.

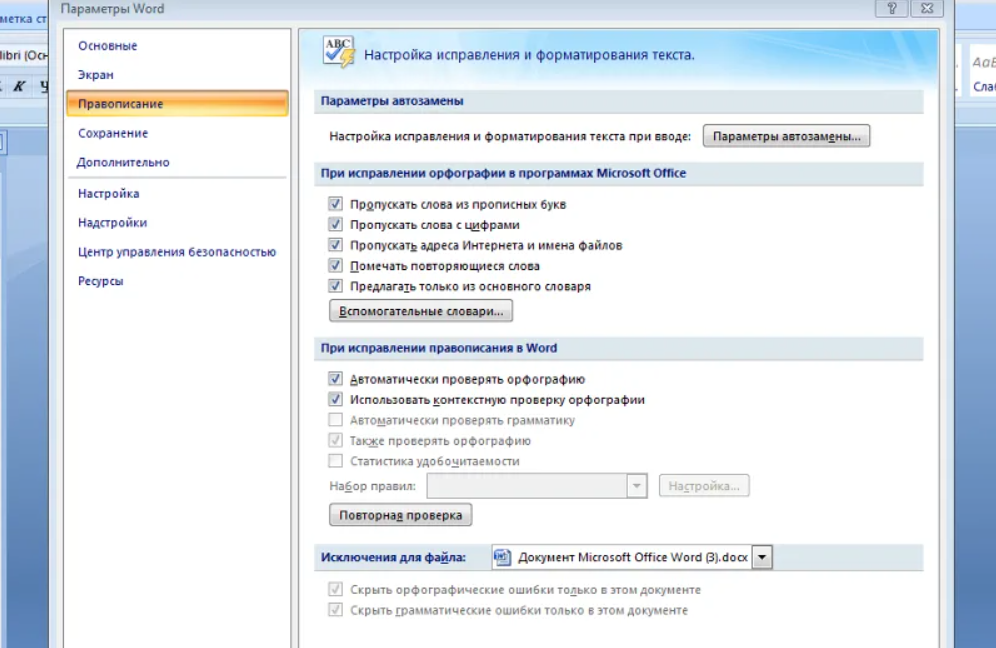


Рисунок 1.3 – Экранная форма настройки исправления и форматирования текста в программе MS Word

Эти инструменты делают процесс проверки текста в Microsoft Word более эффективным и помогают создавать более качественные и профессиональные документы.

Закон Ципфа - это эмпирическое наблюдение, которое было впервые сформулировано лингвистом Джорджем Ципфом в 1935 году. Он обнаружил, что в большинстве языков существует статистическое распределение слов по частоте употребления, и это раcпределение подчиняется математическому закону.

В контексте литературного анализа, закон Ципфа может быть применен к изучению частоты употребления слов в текстах. Суть закона Ципфа в том, что если отсортировать слова в тексте по убыванию их частоты использования, то второе по частоте слово будет использоваться приблизительно в два раза реже, третье — в три раза реже, и так далее. Математически это описывается формулой, где частота (f) обратно пропорциональна рангу (r) слова.

Закон Ципфа имеет важное значение в лингвистике и языковых исследованиях. Он позволяет описывать и моделировать распределение слов в текстах и языках. Этот закон отражает неравномерность использования слов в языке: небольшое количество слов (частотные слова) встречается очень часто, в то время как большинство слов (редкие слова) встречается редко или даже вовсе не встречается.

Закон Ципфа может быть использован для различных лингвистических задач. Например, он может быть применен для оценки уникальности текста, идентификации авторства, классификации текстов, анализа стилей письма и др. Также он может быть использован для построения языковых моделей и прогнозирования вероятности появления слова в тексте.

По сути, закон Ципфа подчеркивает неравномерное распределение частот слов в тексте. Некоторые слова (например, артикли, предлоги) встречаются очень часто, в то время как большинство слов используется гораздо реже. На рисунке 1.4 показано распределение слов в произведении «Война и мир».

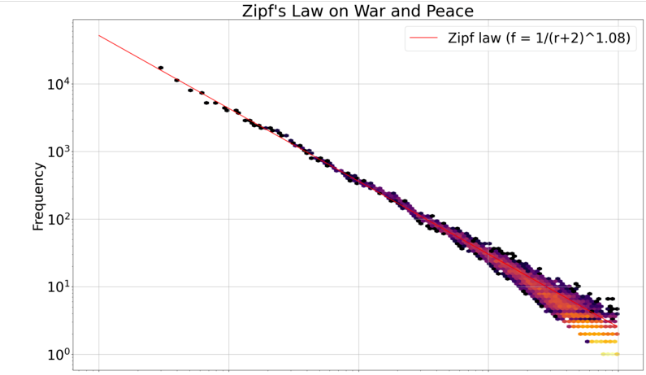


Рисунок 1.4 – График распределения по закону Ципфа на примере «Войны и мира»

В литературном анализе закон Ципфа применяется для изучения структуры текстов, выявления особенностей лексики и выделения наиболее важных слов. Например, в корпусе литературных произведений можно выделить наиболее часто встречающиеся слова, которые могут быть ключевыми для понимания содержания текста. Закон Ципфа также может использоваться для сравнения текстов разных авторов или литературных периодов, что позволяет выявить уникальные стили и особенности использования языка.

Существует также усовершенствованная формулировка закона Ципфа. Развернутый закон Ципфа, также известный как закон Ципфа-Мандельброта, представляет собой более детальное описание статистического распределения слов в тексте или языке.

В развернутом варианте закона добавляется показатель s, который может быть различным для разных текстов и языков. Он отражает степень неравномерности распределения слов. Если s=1, то распределение соответствует оригинальному закону Ципфа, где частота встречаемости слов обратно пропорциональна их рангу. Однако в реальности s обычно отличается от 1 и может принимать значения от 0 до 2.

Значение показателя степени s может быть использовано для характеристики текста или языка. Например, если s близко к 0, это указывает на равномерное распределение слов и низкую концентрацию на частотных словах. Если s близко к 2, это указывает на сильную неравномерность распределения, где небольшое количество слов встречается очень часто.

Применение развернутого закона Ципфа может помочь в следующих задачах:

* анализ текстов: Распределение слов в тексте, описанное развернутым законом Ципфа, может использоваться для анализа и сравнения текстов. Оно может помочь выявить особенности лексического состава текста и сравнить его с другими текстами;
* распознавание языка: Развернутый закон Ципфа может быть использован в задаче распознавания языка для определения языка текста на основе его распределения слов. Различные языки имеют свои уникальные распределения слов, и развернутый закон Ципфа может помочь в их идентификации;
* оценка текстов: Развернутый закон Ципфа может быть использован в задаче оценки качества текста. Анализ распределения слов по рангу и длине может дать представление о структуре и стилистических особенностях текста.

Развернутый закон Ципфа позволяет более точно моделировать и анализировать статистические свойства текстов и языков. Он может быть использован для оценки уникальности текста, идентификации авторства, классификации текстов, анализа стилей письма и других лингвистических задач. Также он может быть использован для построения языковых моделей и прогнозирования вероятности появления слова в тексте.

Важно отметить, что закон Ципфа не всегда строго соблюдается в каждом тексте, и отклонения могут быть обусловлены разными стилями письма, темами текстов или особенностями авторского стиля.

Hemingway Editor — это инструмент, предназначенный для улучшения читаемости текста. Он анализирует текст на сложность, выделяя фрагменты, которые могут быть упрощены, делает акцент на доступности и ясности языка. Инструмент ориентирован на уменьшение сложности предложений и устранение стилистических недостатков.. Экранная форма представлена на рисунке 1.5

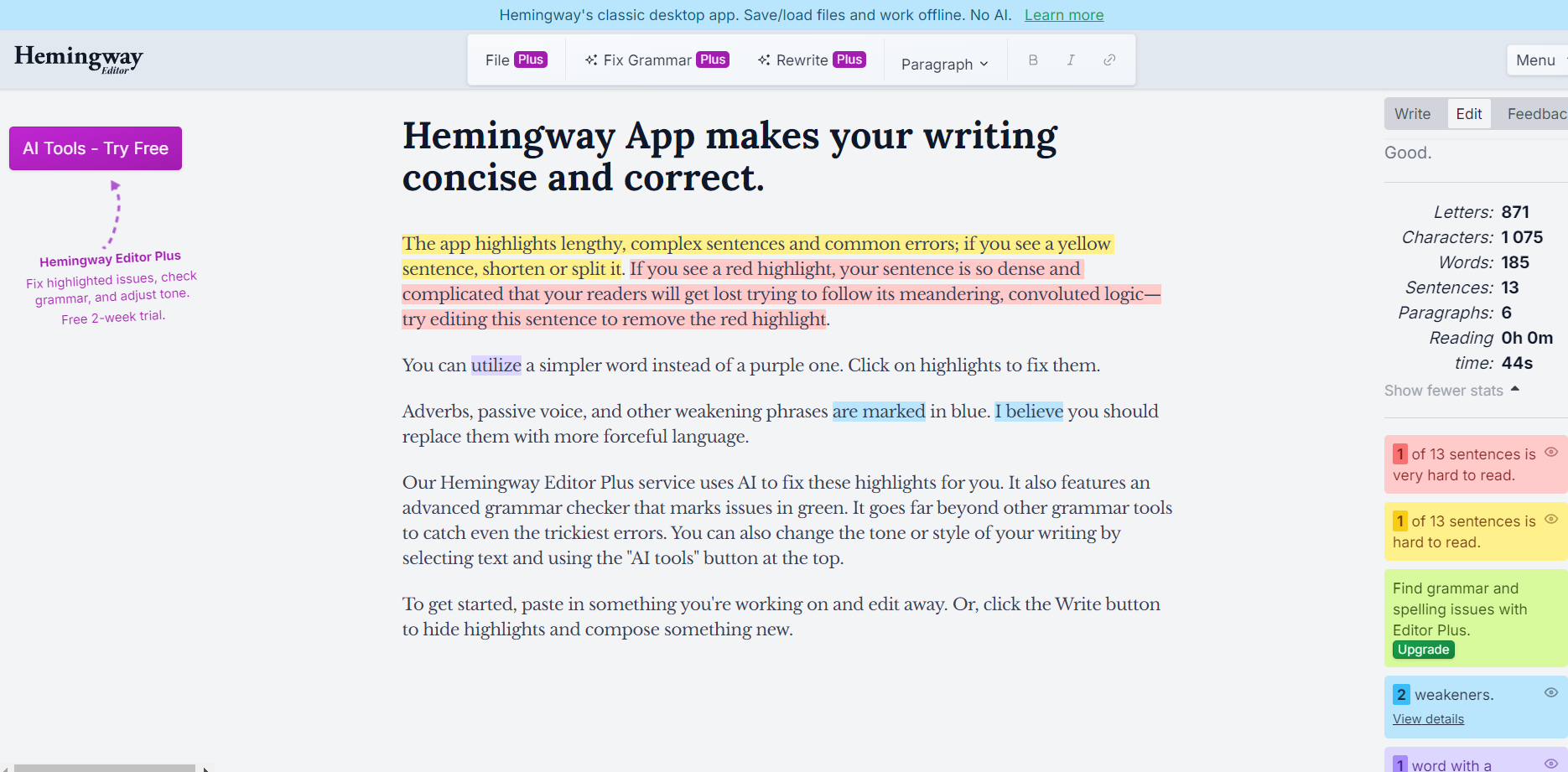


Рисунок 1.5 – Стартовая страница Hemingway Editor

Предлагает советы по улучшению, однако не формирует итоговую оценку, также не поддерживает русский язык.

Еще одним популярным иструментом является ProWritingAid — это инструмент для редактирования текста, ориентированный на профессиональных писателей, редакторов и контент-маркетологов. Он использует более сложные алгоритмы для анализа текста и предлагает исправления на основе многих аспектов, включая стиль, грамматику, синтаксис и структуру текста. При анализе текста, помимо других методов, использует tf-idf – это статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции. Экранная форма представлена на рисунке 1.6.

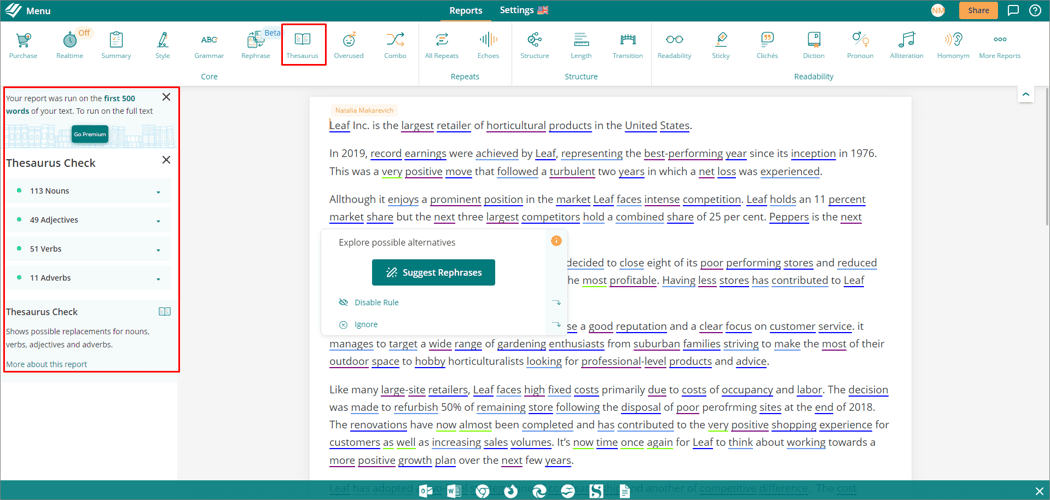


Рисунок 1.6 – Экранная форма ProWritingAid редактирования текста

Имеет широкий функционал, но в то же время перегруженный интерфейс, что затрудняет его использование.

## Обоснование плана диссертационных исследований вообще и методов исследований анализа художественных текстов

Для исследования текста будет применяться многокритериальная оценка.

Многокритериальная оценка (МКО) - это метод, который позволяет принимать решения при наличии нескольких критериев или факторов, которые необходимо учесть. В отличие от однокритериальной оценки, где решение принимается на основе одного критерия, МКО учитывает несколько критериев одновременно.

Процесс многокритериальной оценки обычно состоит из следующих шагов:

* определение целей и критериев: сначала необходимо четко определить цели оценки и критерии, которые будут использоваться для оценки альтернатив. Критерии должны быть измеримыми и релевантными для принимаемого решения;
* сбор данных: для каждой альтернативы необходимо собрать данные, соответствующие каждому из критериев. Данные могут быть числовыми (например, стоимость, время выполнения) или качественными (например, уровень качества, удовлетворенность клиентов);
* взвешивание критериев: после сбора данных необходимо определить относительную важность каждого критерия. Это можно сделать путем присвоения весов каждому критерию, где более важные критерии получают более высокие веса. Веса могут быть определены экспертно или с помощью методов анализа приоритетов, таких как анализ иерархий или анализ парных сравнений;
* нормализация данных: поскольку данные могут иметь различные единицы измерения и диапазоны значений, необходимо нормализовать их перед проведением оценки. Нормализация позволяет привести все критерии к общей шкале и сравнить их между собой;
* ранжирование альтернатив: после нормализации данных можно приступить к ранжированию альтернатив. Существует несколько методов ранжирования, включая взвешенную сумму, взвешенное произведение, метод анализа иерархий и другие. В каждом из методов альтернативы получают оценку на основе весов критериев и их значений;
* принятие решения: после ранжирования альтернатив можно принять решение, выбрав альтернативу с наивысшим рангом или с наиболее высокой оценкой. Однако решение может быть также связано с анализом чувствительности, чтобы оценить, как изменения в весах критериев влияют на результаты.

Метод многокритериальной оценки является полезным инструментом для принятия решений в сложных ситуациях, где необходимо учесть множество факторов. Он позволяет систематизировать процесс оценки и сравнения альтернатив, что помогает принимать обоснованные и информированные решения.

Многокритериальная оценка художественных текстов предполагает использование нескольких критериев или метрик для оценки качества текста. Особенности многокритериальной оценки художественных текстов включают:

* субъективность: Оценка художественных текстов является субъективным процессом, поскольку восприятие и оценка качества текста могут различаться у разных людей. Многокритериальная оценка позволяет учесть различные аспекты текста и учесть разнообразие мнений и предпочтений;
* разнообразие критериев: Многокритериальная оценка включает использование различных критериев или метрик для оценки художественных текстов. Эти критерии могут включать структуру текста, языковые особенности, стиль письма, эмоциональное воздействие, оригинальность и другие аспекты текста;
* взаимодействие критериев: При многокритериальной оценке важно учитывать взаимодействие между различными критериями. Например, критерий структуры текста может влиять на критерий языковых особенностей, и оба могут влиять на критерий эмоционального воздействия. Важно учесть эти взаимосвязи при оценке текста;
* весовые коэффициенты: В многокритериальной оценке каждый критерий может иметь различную важность или вес. Весовые коэффициенты определяют, насколько каждый критерий влияет на общую оценку текста. Определение этих весовых коэффициентов может быть сложной задачей и требует внимательного анализа и экспертных оценок;
* агрегация результатов: После оценки текста по каждому критерию необходимо агрегировать результаты для получения общей оценки. Существуют различные методы агрегации, такие как суммирование, усреднение, взвешенное суммирование и другие. Выбор метода агрегации будет зависеть от особенностей конкретного исследования;
* интерпретация результатов: После проведения многокритериальной оценки важно интерпретировать полученные результаты. Это может включать анализ важности разных критериев, выявление особенностей текста, идентификацию сильных и слабых сторон и общую оценку качества текста.

Многокритериальная оценка художественных текстов является сложным и многогранным процессом, который требует внимательного анализа и экспертных знаний. Она позволяет учесть разнообразие аспектов текста и получить оценку его качества, основываясь на большем количестве критериев.

В таблице 1.1 представлена сводная информация о существующих методах анализа художественных текстов.

Таблица 1.1 – Анализ предметной области

|  |  |
| --- | --- |
| Название ПО | Используемый метод |
| ProWritingAid | TF-IDF  , где tf - частота слова, idf – обратная частота |
| Hemingway Editor |  |
| Лингвистический анализ «Лаборатории фантастики» | Удельный авторский словарный запас (АСЗ) |

Существующие решения имеют следующие недостатки:

* отсутствие возможности настройки уровня влияния характеристик;
* игнорирование контекста при вычислении численных показателей;
* отсутствие итоговой оценки;
* игнорирование низкоуровневых критериев оценки.

Для устранения этих недостатков разрабатывается новая методика.

## Обоснование пути решения проблемы разработки алгоритма и методики художественных текстов

Для устранения описанных недостатков производится разработка новых методики и алгоритма многокритериальной оценки художественных текстов. Для решения поставленных задач предлагается использовать нейронные сети. Оценка художественных текстов с использованием нейронных сетей представляет собой один из современных подходов к решению этой проблемы. Нейронные сети - это алгоритмы машинного обучения, которые моделируют работу нейронной системы человека и способны обрабатывать сложные данные, включая тексты.

Процесс оценки художественных текстов с использованием нейронных сетей может быть следующим:

* подготовка данных: Сначала необходимо подготовить данные, на которых будет обучаться нейронная сеть. Это может включать сбор и разметку текстовых данных, которые будут использоваться для обучения и тестирования модели;
* создание модели нейронной сети: затем необходимо создать архитектуру нейронной сети, которая будет использоваться для оценки художественных текстов. Это может быть рекуррентная нейронная сеть (RNN), сверточная нейронная сеть (CNN) или комбинация различных типов слоев;
* обучение модели: после создания модели необходимо обучить ее на подготовленных данных. Это включает процесс передачи текстовых данных через сеть и корректировку весов нейронов на основе задачи оценки художественных текстов;
* валидация и тестирование модели: после обучения модели необходимо провести валидацию и тестирование для оценки ее производительности. Это может включать использование отложенного набора данных для проверки точности и обобщающей способности модели;
* оценка текстов: после успешного обучения и проверки модели можно использовать ее для оценки художественных текстов. Это может быть процессом классификации текстов на основе определенных категорий или предсказанием числовых оценок, связанных с качеством текста.

Преимущества использования нейронных сетей для оценки художественных текстов включают способность моделировать сложные зависимости в тексте, автоматическую обработку большого объема данных и возможность обучения на большом количестве разнообразных текстов. Однако важно отметить, что эффективность и точность модели зависят от качества подготовленных данных, выбранной архитектуры сети и правильного обучения и настройки модели.

Оценка художественных текстов с использованием нейронных сетей имеет несколько преимуществ:

* способность моделировать сложные зависимости: Нейронные сети обладают способностью обрабатывать и моделировать сложные зависимости в тексте. Это позволяет им улавливать нюансы и контекстуальные особенности художественных текстов, которые могут влиять на их оценку. Нейронные сети могут автоматически извлекать признаки из текста и использовать их для принятия решений об оценке;
* автоматическая обработка большого объема данных: Нейронные сети могут эффективно обрабатывать большой объем текстовых данных. Это позволяет им изучать широкий спектр художественных текстов и учитывать разнообразие стилей, жанров и тематик. Большой объем данных также способствует более точному обучению модели и улучшению ее способности оценивать тексты;
* учет контекста и субъективности: Нейронные сети могут быть настроены на учет контекста и субъективности при оценке художественных текстов. Они могут учитывать особенности авторского стиля, использование литературных приемов и эмоциональную окраску текста. Это позволяет более точно оценивать качество и ценность текста с учетом его специфических характеристик;
* использование объективных метрик: Нейронные сети могут быть обучены на основе объективных метрик, которые помогают измерить различные аспекты художественных текстов, такие как сложность сюжета, грамматическая правильность, использование литературных приемов и другие. Это позволяет более объективно оценивать тексты и учитывать различные критерии при оценке;
* развитие новых критериев оценки: Оценка художественных текстов с использованием нейронных сетей может способствовать развитию новых критериев оценки. Нейронные сети могут обучаться на основе различных данных, включая оценки экспертов или предпочтения читателей. Это может помочь выявить новые аспекты, которые влияют на качество текста и улучшить процесс оценки.

Однако важно отметить, что эффективность и точность оценки художественных текстов с использованием нейронных сетей зависят от качества подготовленных данных, выбранной архитектуры сети и правильного обучения и настройки модели. Это требует тщательного исследования и экспериментов для достижения наилучших результатов.

## Выводы по главе 1

Исследование показало, что существуют различные методы анализа художественных текстов. Некоторые из них включают в себя анализ структуры и синтаксических конструкций текста, изучение лексики и ее влияние на смысл произведения, анализ образов и метафор, исследование грамматических особенностей и стилистических приемов.

Также было обнаружено, что использование подходов, основанных на компьютерной обработке текстов, например, с использованием методов машинного обучения и анализа данных, позволяет проводить более глубокий и объективный анализ художественных текстов.

Однако, несмотря на разнообразие методов анализа текстов, было отмечено, что ни один из них не может полностью заменить традиционные методы анализа, такие как чтение и интерпретация текстов человеком. Таким образом, использование разнообразных методов, включая традиционные методы и новые технологии, можно добиться более полного и всестороннего понимания художественных текстов.

# ГЛАВА 2. ФОРМАЛИЗОВАННОЕ ПРЕДСТАВЛЕНИЕ ЗАДАЧИ АНАЛИЗА ХУДОЖЕСТВЕННЫХ ТЕКСТОВ

## Разбиение проблемы на подпроблемы. Формальные способы и пути решения подпроблем анализа художественных текстов

Автоматизированный алгоритм многокритериальной оценки художественных текстов представляет собой сложный процесс, который включает в себя несколько этапов:

* подготовка данных: в начале необходимо подготовить набор данных, который будет использоваться для оценки. Этот набор данных должен содержать тексты, которые будут оцениваться, а также соответствующие метки или оценки для каждого из установленных критериев. Каждый текст может быть представлен в виде последовательности слов или в виде другого текстового представления;
* выбор критериев оценки: определите критерии, которые будут использоваться для оценки текстов. Эти критерии могут включать в себя лексическое богатство, грамматическую правильность, синтаксическую сложность, семантическую связность, эмоциональную окраску, оригинальность, художественную ценность и тематическую разнообразность;
* предварительная обработка текста: тексты могут потребовать предварительной обработки, такой как удаление стоп-слов, приведение к нижнему регистру, токенизация и лемматизация;
* извлечение признаков: для каждого текста извлекаются признаки, соответствующие выбранным критериям. Например, для оценки лексического богатства можно вычислить количество уникальных слов в тексте. Для оценки грамматической правильности можно использовать статистику по частоте ошибок в тексте;
* моделирование: далее строится модель, которая принимает на вход признаки текстов и предсказывает значения для каждого из установленных критериев. В зависимости от выбранного метода могут использоваться различные модели машинного обучения, такие как линейная регрессия, случайный лес, нейронные сети и т.д.;
* оценка: полученные предсказания модели сравниваются с истинными значениями критериев для оценки текстов. Можно использовать различные метрики для оценки качества модели, такие как среднеквадратичная ошибка, коэффициент детерминации, точность, полнота и т.д.;
* интерпретация результатов: полученные результаты анализируются для выявления тенденций и понимания того, какие аспекты текста оказывают наибольшее влияние на его качество. Это может помочь авторам или редакторам литературных произведений лучше понимать сильные и слабые стороны текстов и улучшать их качество;
* доработка и улучшение: на основе анализа результатов модель и процесс оценки могут быть доработаны и улучшены для получения более точных и надежных оценок качества художественных текстов.

Для формализации задачи многокритериальной оценки художественных текстов необходимо определить критерии, по которым будет осуществляться оценка. Среди критериев выделены следующие:

* активный словарный запас - это набор слов, которые автор регулярно использует в своей речи и письме. Это слова, которые автор может легко вспомнить и использовать для выражения своих мыслей и идей. Активный словарный запас обычно меньше, чем пассивный словарный запас, который включает в себя слова, которые автор понимает, но не использует регулярно;
* удельный авторский словарный запас;
* доля диалогов в тексте – определяется как число авторских знаков, участвующих в диалогах, к общему количеству авторских знаков в исследуемом тексте;
* средняя длина предложения;
* статистика использования частей речи - процент существительных, прилагательных, глаголов, местоимений-существительных, местоименных прилагательных, местоимений-предикативов, числительных и др;
* удельное количество ошибок – в данном критерии оценивается относительная частота орфографических и пунктуационных ошибок.

Удельный авторский словарный запас можно вычислить с помощью различных методов, однако одним из наиболее распространенных и простых является использование индекса лексического богатства текста, такого как индекс Херша (Herfindahl index) или индекс Шеннона (Shannon index).

Индекс Херша вычисляется путем подсчета относительной частоты употребления каждого слова в тексте и последующего вычисления суммы квадратов этих относительных частот. Формула для вычисления индекса Херша выглядит следующим образом:

|  |  |
| --- | --- |
|  | (2.1) |

- относительная частота употребления i-го слова, а

n - общее количество слов в тексте.

Индекс Шеннона (Shannon index):

Индекс Шеннона используется для измерения степени разнообразия или неопределенности в наборе данных, в данном случае, в тексте. Он также учитывает как частоту, так и равномерность распределения слов в тексте. Формула для вычисления индекса Шеннона выглядит следующим образом:

|  |  |
| --- | --- |
|  | (2.2) |

- вероятность появления i-го слова в тексте (относительная частота), а

n - общее количество слов в тексте.

После вычисления индексов, таких как индекс Херша или Шеннона, можно заключить, что их более высокие значения свидетельствуют о большем разнообразии и богатстве словарного запаса текста. Для количественной оценки удельного авторского словарного запаса предлагается нормализовать значение индекса, разделив его на общее число слов в тексте.

Эти методы обеспечивают обобщённую оценку лексического разнообразия, что делает их эффективными инструментами для анализа и сравнения словарного богатства различных текстов, а также для наблюдения за его изменениями с течением времени. Тем не менее, важно учитывать, что такие подходы не отражают семантического содержания слов, что может играть существенную роль в изучении индивидуального стиля автора.

В дополнение к этому, исследование статистических характеристик, связанных с частотой употребления частей речи в литературных произведениях, является значимым этапом анализа их языковых особенностей. Этот процесс может быть разделен на несколько этапов:

* токенизация и разметка частей речи: Первый шаг - разбиение текста на отдельные слова или токены и определение частей речи для каждого токена. Для этого используются методы естественного языка и инструменты, такие как NLTK (Natural Language Toolkit) или SpaCy. Разметка частей речи позволяет идентифицировать глаголы, существительные, прилагательные, наречия и другие части речи в тексте.
* вычисление частотности: Для каждой части речи вычисляется ее частотность в тексте. Это делается путем подсчета количества вхождений каждой части речи и деления этого числа на общее количество слов или токенов в тексте. Например, можно вычислить процентное соотношение глаголов, существительных и других частей речи относительно общего числа слов в тексте.
* анализ структуры текста: Полученная статистика позволяет проанализировать структуру текста и выявить его особенности. Например, высокая частота глаголов может указывать на динамичный сюжет, а высокая частота существительных может свидетельствовать о детальном описании окружающего мира или персонажей.
* сравнение с нормативными данными: Для более глубокого анализа можно сравнить полученную статистику с нормативными данными или данными других текстов. Например, можно сравнить частотность использования глаголов в художественных текстах с их частотностью в научных статьях или новостных материалах, чтобы определить уникальные особенности стиля художественных произведений.

Начальным этапом сбора данных является создание обучающего набора, который включает размеченные художественные тексты с заранее известными значениями всех критериев оценки. Этот набор должен быть тщательно сформирован с учётом его разнообразия и репрезентативности, чтобы обеспечить высокое качество обучения модели. Для формирования базы данных используются тексты, добровольно опубликованные авторами в открытых источниках. Сбор данных из подобных ресурсов позволяет сформировать обширный и качественный набор, который затем применяется для анализа и выработки стандартных значений по установленным критериям.

Автоматизированная оценка текстов основывается на использовании алгоритмов и методов машинного обучения, способных эффективно обрабатывать текстовые данные и рассчитывать значения для каждого критерия. Эти методы варьируются от классических алгоритмов анализа текстов до более продвинутых подходов глубокого обучения, таких как нейронные сети.

Одной из ключевых стадий анализа данных является нормализация. Она имеет важное значение в многокритериальной оценке, так как позволяет сопоставлять параметры, различающиеся по своим единицам измерения, диапазонам значений и природе. Нормализация обеспечивает единообразие и корректность анализа, что крайне важно при сравнении различных альтернатив и интерпретации результатов.

Нормализация производится со следующими целями:

* сделать данные сопоставимыми, устранив различия в единицах измерения и диапазонах значений;
* преобразовать данные в единый диапазон, обычно от 0 до 1.

Существует несколько методов нормализации. Примеры методов нормализации:

1. Линейная нормализация

* формула: X' = (X - Xmin) / (Xmax - Xmin); (2.3)
* преобразует данные в диапазон от 0 до 1;
* подходит для критериев, где более высокие значения предпочтительнее.

2. Нормализация по сумме

* формула: X' = X / ΣX; (2.4)
* преобразует данные в пропорции, где сумма всех нормализованных значений равна 1;
* подходит для критериев, где значения должны складываться до 100%.

3. Нормализация по максимуму

* формула: X' = X / Xmax; (2.5)
* преобразует данные в диапазон от 0 до 1, где максимальное значение равно 1;
* подходит для критериев, где более высокие значения предпочтительнее.

4. Нормализация по минимуму

* формула: X' = (Xmin - X) / (Xmin - Xmax); (2.6)
* преобразует данные в диапазон от 0 до 1, где минимальное значение равно 1;
* подходит для критериев, где более низкие значения предпочтительнее.

5. Нормализация по рангу

* присваивает каждому значению ранг от 1 (лучшее) до n (худшее);
* нормализованное значение рассчитывается как ранг / n.

После выполнения нормализации данных возможно использование различных методов, таких как взвешенные суммы или другие подходы, для определения окончательного рейтинга альтернатив. Выбор подходящего метода нормализации зависит от специфики задачи и целей многокритериальной оценки.

Агрегирование критериев представляет собой процесс интеграции нескольких отдельных критериев в единый итоговый показатель, характеризующий совокупную оценку альтернатив. Для этого применяются различные методы, каждый из которых имеет свои сильные и слабые стороны.

Одним из наиболее популярных подходов к агрегированию является метод взвешенной суммы. Его суть заключается в суммировании произведений нормализованных значений критериев на их соответствующие веса, которые отражают степень значимости каждого критерия. Этот метод позволяет получить итоговую оценку, учитывающую относительную важность различных факторов, и широко используется благодаря своей простоте и эффективности. Существует два основных типа методов взвешенной суммы:

* линейная взвешенная сумма (LWS): альтернативы оцениваются путем суммирования произведений значений критериев на их веса.
* экспоненциальная взвешенная сумма (EWS): похожа на LWS, но веса возводятся в степень, которая определяет относительную важность критериев.

Преимущества методов взвешенной суммы:

* простота в понимании и реализации;
* возможность учета относительной важности критериев;
* возможность использования различных шкал измерения для разных критериев.

Недостатки методов взвешенной суммы:

* зависимость от точности весов;
* трудность в определении весов, особенно когда критерии имеют разные единицы измерения;
* возможность компенсации между критериями (т.е. низкое значение по одному критерию может быть компенсировано высоким значением по другому критерию).

Применение методов взвешенной суммы: методы взвешенной суммы широко используются в различных областях, таких как:

* выбор поставщиков;
* оценка инвестиционных проектов;
* ранжирование кандидатов на работу;
* анализ рисков.

Выбор между методами LWS и EWS определяется спецификой задачи многокритериальной оценки. Метод LWS характеризуется простотой реализации и высокой эффективностью при наличии линейных зависимостей между критериями. В свою очередь, метод EWS предоставляет возможность учитывать сложные нелинейные взаимодействия между критериями, что делает его более универсальным инструментом для решения задач повышенной сложности.

Методы лексикографического порядка являются подходом к агрегированию критериев, при котором альтернативы упорядочиваются последовательно на основе одного критерия за другим, начиная с наиболее приоритетного. Такой подход позволяет учитывать различия в значимости критериев и обеспечивает ранжирование альтернатив с акцентом на ключевые показатели. Методы лексикографического порядка:

* метод Максимина: выбирается альтернатива с наилучшим значением по самому важному критерию. Если несколько альтернатив имеют одинаковое значение, переходят к следующему по важности критерию;
* метод Максикола: похож на метод Максимина, но выбирается альтернатива с наилучшим значением по наименее важному критерию.

Преимущества методов лексикографического порядка:

* гарантируют, что наиболее важные критерии имеют приоритет в процессе принятия решения;
* просты в понимании и реализации;
* не требуют определения весов для критериев.

Недостатки методов лексикографического порядка:

* могут привести к нежелательным результатам, когда значения критериев близки друг к другу;
* не учитывают компенсацию между критериями;
* могут быть неэффективны, когда количество критериев велико.

Методы лексикографического порядка часто используются в ситуациях, когда:

* критерии имеют четкую иерархию важности;
* необходимо избежать компенсации между критериями;
* количество критериев относительно невелико.

Методы попарного сравнения - это методы принятия решений, которые включают сравнение каждой пары альтернатив и выбор предпочтительной альтернативы из каждой пары. Эти методы часто используются, когда имеется несколько альтернатив и необходимо выбрать лучшую. Существующие методы:

* метод аналитической иерархии процесса (AHP): Критерии сравниваются попарно, и их относительная важность определяется с помощью шкал предпочтений;
* метод предпочтения (PROMETHEE): Критерии сравниваются попарно, и для каждой пары рассчитывается положительное и отрицательное отклонение. Затем отклонения агрегируются для получения общего индекса предпочтения.

Методы попарного сравнения имеют ряд преимуществ, в том числе:

* простота использования;
* возможность учитывать несколько критериев;
* возможность учитывать неопределенность и предпочтения.

Недостатки методов попарного сравнения есть и ряд недостатков, в том числе:

* они могут быть трудоемкими, особенно когда имеется большое количество альтернатив;
* они могут быть подвержены ошибкам, особенно когда сравниваются альтернативы, которые очень похожи;
* они могут не всегда давать последовательные результаты.

Другие методы:

* метод TOPSIS (Technique for Order Preference by Similarity to Ideal Solution): Альтернативы оцениваются на основе их расстояния от идеального и отрицательного идеального решений;
* метод VIKOR (VlseKriterijumska Optimizacija I Kompromisno Resenje): Альтернативы ранжируются по двум критериям: расстоянию от идеального решения и расстоянию от отрицательного идеального решения;
* метод ELECTRE (Elimination and Choice Expressing Reality): Альтернативы оцениваются на основе их согласованности и несогласованности с критериями.

Выбор метода агрегирования:

Выбор подходящего метода агрегирования зависит от конкретной задачи оценки. Факторы, которые следует учитывать, включают:

* количество критериев;
* относительная важность критериев;
* наличие данных;
* предпочтения лиц, принимающих решения.

Необходимо отметить, что ни один из методов агрегирования критериев не является абсолютно идеальным. Итоговые результаты оценки могут существенно зависеть от выбранного метода. Поэтому целесообразно использовать несколько подходов и сравнивать полученные результаты, что способствует повышению точности и обоснованности анализа.

В рамках многокритериальной оценки лексикографическое отношение предпочтения представляет собой метод принятия решений, при котором критерии упорядочиваются по степени их важности. Сравнение альтернатив осуществляется, начиная с самого значимого критерия. Если по этому критерию альтернативы оказываются равнозначными, анализ переходит к следующему критерию в порядке убывания его значимости. Процесс продолжается до момента обнаружения различий или завершения сравнения по всем критериям. Такой метод гарантирует строгую последовательность учёта значимости факторов, что позволяет эффективно принимать решения в условиях многокритериальной оценки. Данное отношение строится по следующему алгоритму:

* упорядочение критериев: критерии упорядочиваются от наиболее важного к наименее важному;
* сравнение по наиболее важному критерию: альтернативы сравниваются по наиболее важному критерию; если альтернатива превосходит другую по этому критерию, ей отдается предпочтение;
* при равенстве по наиболее важному критерию - если альтернативы равны по наиболее важному критерию, они сравниваются по второму наиболее важному критерию;
* повторение процесса - процесс повторяется для каждого критерия в порядке важности, пока не будет установлено предпочтение.

Преимущества:

* простота и легкость применения:
* обеспечивает четкую иерархию критериев;
* позволяет принимать решения даже при наличии противоречивых критериев.

Недостатки:

* метод может быть проблематичным, если критерии не могут быть четко упорядочены;
* может привести к игнорированию важных критериев, если они не являются наиболее важными;
* не учитывает относительную важность критериев.

На основе этого метода была разработана его улучшенная версия - метод последовательных уступок.

Метод последовательных уступок — это метод многокритериальной оценки, который используется для определения относительной важности различных критериев и поиска решений, которые удовлетворяют нескольким конфликтующим критериям.

Процесс метода последовательных уступок:

* определение критериев и их важности: необходимо определить все релевантные критерии, которые будут использоваться для оценки решений, после чего присвоить каждому критерию вес, отражающий его относительную важность;
* ранжирование решений: требуется оценить каждое решение по каждому критерию и ранжируйте их в порядке убывания желательности;
* выявление доминирующих решений: на данном этапе определяются решения, которые доминируют над другими по всем критериям. Если такое решение существует, оно является предпочтительным;
* последовательные уступки: если нет доминирующих решений, начните с самого важного критерия, в ином случае рассматриваются решения, которые имеют наилучший рейтинг по этому критерию; если несколько решений имеют одинаковый рейтинг, необходимо перейти к следующему по важности критерию и сделайте уступки по первому критерию. Процесс необходимо продолжать, пока не будет достигнуто решение, которое удовлетворяет всем критериям в приемлемой степени.

Преимущества метода последовательных уступок:

* простота: метод относительно прост в понимании и применении;
* учет нескольких критериев: он позволяет учитывать несколько конфликтующих критериев при принятии решений;
* гибкость: метод можно адаптировать к различным проблемам и критериям;
* прозрачность: процесс принятия решений прозрачен и может быть прослежен.

Недостатки метода последовательных уступок:

* субъективность: веса критериев и оценки решений могут быть субъективными;
* компромисс: метод может привести к компромиссным решениям, которые не являются оптимальными по всем критериям;
* сложность: при большом количестве критериев или решений процесс может стать сложным.

В рамках многокритериальной оценки планируется использовать метод лексикографического отношения порядка. Ключевым элементом данного подхода является возможность динамического формирования ранжированного списка критериев, который определяется пользователем. Это позволяет изменять порядок критериев в зависимости от предпочтений и приоритетов, заданных пользователем, что делает процесс оценки более гибким и персонализированным.

Для каждого критерия необходимо проводить расчёт значений различных параметров, которые описывают его характеристики. Кроме того, каждому параметру следует назначить весовой коэффициент, отражающий его значимость и степень влияния на итоговый результат. Итоговое значение критерия рассчитывается с учётом этих весовых коэффициентов, что обеспечивает точное моделирование относительной важности параметров. Такой подход способствует адаптации процесса оценки к специфическим особенностям задачи, позволяет учитывать индивидуальные предпочтения пользователя и предоставляет возможность гибко управлять влиянием различных параметров на общий результат. Значение по каждому критерию с учетом весового коэффициента может быть вычислено по следующей формуле:

|  |  |
| --- | --- |
|  | (2.7) |

Где:

- значимость (весовой коэффициент) критерия оценки текста;

- значение выбранного критерия.

После вычисления значения параметра для каждого из критериев, необходимо вычислить итоговую оценку. Итоговая оценка вычисляется как сумма весовых параметров для каждого из критериев. Расчет итоговой оценки производится по следующей формуле:

|  |  |
| --- | --- |
|  | (2.8) |

Где:

Q – итоговая оценка текста;

– число критериев оценки;

- значимость i-го критерия оценки текста;

- значение i-го критерия оценки текста.

После завершения обучения модель можно протестировать на тестовом наборе данных, чтобы оценить ее производительность и точность. Это позволяет проверить, насколько хорошо модель способна оценивать новые тексты, которые не входили в обучающий набор данных.

Многослойный перцептрон (MLP) представляет собой классическую модель искусственной нейронной сети, основанную на прямом распространении информации без циклических связей. В ее основе лежит идея последовательного преобразования входных данных через один или несколько скрытых слоев, что позволяет аппроксимировать сложные нелинейные зависимости.

Формально, пусть x– вектор входных данных. Модель MLP состоит из L слоев, где первый слой – входной, последний – выходной, а между ними находятся скрытые слои. Каждый слой i характеризуется весовой матрицей Wi и вектором смещений bi.

Здесь f — функции активации, такие как сигмоида, ReLU или tanh, которые вводят нелинейность в модель и позволяют сети решать задачи, выходящие за рамки линейной регрессии.

Процесс обучения MLP обычно реализуется посредством обратного распространения ошибки (backpropagation) в сочетании с методами оптимизации (например, стохастическим градиентным спуском или алгоритмом Adam). Суть метода состоит в следующем: сначала вычисляется разница между предсказанным выходом и истинным значением через выбранную функцию потерь (например, среднеквадратичная ошибка для регрессии), затем эта ошибка «распространяется» назад по сети для корректировки весов Wi и смещений bi с целью минимизации функции потерь.

Одним из важных теоретических результатов является так называемая теорема о универсальном аппроксиматоре, которая утверждает, что однослойный перцептрон (при наличии достаточного числа нейронов и подходящей функции активации) может аппроксимировать любую непрерывную функцию на компактном множестве. Добавление скрытых слоев позволяет не только улучшить аппроксимационную способность модели, но и сократить количество нейронов, необходимых для достижения требуемой точности.

Кроме того, для повышения обобщающей способности сети и предотвращения переобучения широко применяются техники регуляризации, такие как L2-регуляризация, dropout и нормализация батча (batch normalization).

Таким образом, Многослойный перцептрон является мощным инструментом для решения задач регрессии и классификации, обеспечивая возможность моделирования сложных зависимостей посредством комбинации линейных преобразований и нелинейных активаций, а процесс его обучения основан на эффективном алгоритме обратного распространения ошибки, позволяющем адаптивно корректировать параметры модели.

## Показатели эффективности многокритериальной оценки художественных текстов

Лексикографический метод многокритериальной оптимизации основывается на принципе упорядочивания критериев по степени их значимости. В рамках этого метода критерии анализируются последовательно, начиная с наиболее важного. Остальные критерии принимаются во внимание только в тех случаях, когда по предыдущим, более приоритетным критериям невозможно установить различие между альтернативами.

Главная задача оптимизации в данном подходе заключается в нахождении такого порядка критериев, который обеспечивает достижение наилучшего результата в соответствии с поставленными целями. Это делает лексикографический метод особенно полезным в ситуациях, требующих строгого соблюдения приоритетов. Применение данного метода обеспечивает рациональность и обоснованность процесса принятия решений, позволяя учитывать ключевые аспекты задачи.

Основные шаги оптимизации лексикографического метода многокритериальной оценки:

* определение критериев: сначала определяются все критерии, которые будут использоваться для оценки альтернативных решений;
* упорядочение критериев: критерии упорядочиваются по их значимости или приоритету. Это может быть сделано на основе экспертных оценок, статистических данных, предпочтений заинтересованных сторон или других методов;
* поиск оптимальных решений: затем анализируются альтернативные решения в порядке упорядочения критериев. Для каждого критерия сравниваются значения для различных альтернатив, и выбирается лучшая альтернатива в соответствии с этим критерием;
* проверка на согласованность: проверяется согласованность полученных решений с заданным порядком критериев. Если альтернатива предпочтительна по более важному критерию, но неудовлетворительна по менее важному критерию, это может потребовать пересмотра решения;
* адаптация порядка критериев: в случае необходимости порядок критериев может быть адаптирован на основе анализа результатов и предпочтений принимающего решение;
* итерационный процесс: процесс поиска оптимальных решений может быть итеративным, особенно если требуется достижение компромисса между различными аспектами принимаемых решений;
* оценка результатов: наконец, производится оценка полученных результатов с учетом поставленных целей и требований задачи. Это позволяет определить, насколько удовлетворительными являются найденные оптимальные решения.

Оптимизация лексикографического метода многокритериальной оценки предполагает глубокий анализ критериев, их взаимной значимости и характера взаимодействия. Важной составляющей является обеспечение гибкости подхода, что позволяет учитывать изменяющиеся условия и предпочтения лица, принимающего решения. Такой подход интегрирует объективные характеристики и субъективные предпочтения, способствуя повышению точности и актуальности итоговых выводов.

Для проверки эффективности результатов, полученных в ходе анализа, будет использоваться метрика, основанная на разнице между рассчитанной ошибкой и так называемой «экспертной» оценкой. В качестве экспертных данных предполагается привлекать оценки пользователей, опубликованные на специализированных онлайн-платформах. Сопоставление результатов автоматизированного анализа с пользовательскими оценками позволит объективно оценить точность модели и её соответствие реальным ожиданиям в практических сценариях.Таким образом, при ранжировании критериев для многокритериальной оценки с использованием отношения лексикографического порядка будет использоваться отношение:

|  |  |
| --- | --- |
|  | (2.9) |

Где:

– оценка текста, полученная согласно методу многокритериальной оценки с отношением лексикографического порядка;

– экспертная оценка, сформированная пользователями специализированных интернет-ресурсов.

Данное соотношение будет использоваться для оценки эффективности многокритериальной оценки.

В качестве эталонной оценки может быть использовано как сравнение с экспертными оценками, так и сравнение с читательскими рейтингами.

Экспертные оценки обычно представляют собой субъективные мнения профессионалов, таких как литературные критики, редакторы или преподаватели литературы. Эти специалисты обладают обширным опытом и глубокими знаниями в области анализа художественных текстов, что позволяет им давать детализированные и обоснованные оценки произведений. Их суждения часто воспринимаются как эталонные, поскольку они опираются на профессиональное понимание литературных стандартов, стилевых особенностей и художественной ценности текста. Применение таких оценок в качестве контрольных данных служит важным инструментом для проверки точности и надёжности автоматизированных методов анализа, обеспечивая высокий уровень доверия к полученным результатам.

Чтобы сравнить алгоритм с оценками экспертов, можно использовать следующие шаги:

1. сбор данных: необходимо собрать корпус текстов, для которых уже существуют оценки экспертов. Оценки могут быть в форме числовых баллов (например, по шкале от 1 до 10) или категориальных значений (например, "плохой", "хороший", "отличный"). После чего следует собрать мнения от нескольких экспертов, чтобы учесть субъективность человеческих оценок. Среднее или медианное значение оценок поможет уменьшить влияние индивидуальных предпочтений;
2. запуск алгоритма: следует применить алгоритм к каждому тексту из корпуса и получить оценки на основе формальных признаков (например, словарного запаса, длины предложений и прочего);
3. корреляционный анализ: определение корреляции между оценками, полученными с помощью алгоритма и с помощью экспертов. Для этой цели можно использовать коэффициент корреляции Пирсона (Используется для оценки линейной зависимости между числовыми оценками экспертов и оценками вашего алгоритма) или коэффициент корреляции Спирмена (предпочтителен, если данные нельзя считать линейно распределенными или если экспертные оценки имеют ранговую природу. Он показывает, насколько согласованы ранжирования текстов по алгоритму и по мнению экспертов). Выбор между этими методами зависит от характера данных. Коэффициент Пирсона применяется для анализа линейных зависимостей, тогда как коэффициент Спирмена более универсален в случаях, когда оценки имеют ранговую природу или данные не подчиняются линейным закономерностям;
4. анализ отклонений: оцените среднеквадратичное отклонение (MSE) или среднюю абсолютную ошибку (MAE) между оценками экспертов и оценками алгоритма. Это позволит понять, насколько близки оценки алгоритма к экспертным и где находятся наиболее значимые расхождения;
5. визуализация результатов: постройте графики рассеяния (scatter plot), где по одной оси будут экспертные оценки, а по другой — оценки алгоритма. Это поможет наглядно увидеть зависимость и выявить тексты, в которых мнения экспертов и алгоритма сильно расходятся;
6. кросс-валидация на разных экспертах: данный шаг является опциональным и заключается в том, чтобы использовать оценки разных экспертов как тестовые и обучающие данные. Так, часть оценок можно использовать для настройки параметров модели, а затем проверить её предсказательную способность на других экспертных оценках.

Альтернативным подходом для анализа является использование читательских рейтингов. Такие рейтинги представляют собой оценки и отзывы, которые оставляют обычные читатели на литературных платформах и в онлайн-сообществах. Преимущество читательских рейтингов заключается в том, что они отражают мнения широкой и разнообразной аудитории, включающей людей с различным опытом, вкусами и предпочтениями. Сравнение результатов алгоритма с читательскими рейтингами позволяет учитывать массовое восприятие текстов, что делает анализ более универсальным. Этот подход также может выявить различия между профессиональной экспертной оценкой и общественным мнением, что особенно важно для изучения факторов, влияющих на популярность литературных произведений.Алгоритм сравнения тогда будет состоять из следующих шагов:

1. сбор данных: следует начать с того, чтобы собрать читательские рейтинги для большого числа текстов, которые могут быть оценены алгоритмом. Оценки могут быть в виде чисел (звездочек) или текста (отзывы), затем оценить и извлечь из текстовых отзывов ключевые слова и тематические аспекты (например, упоминания о стиле, интересности сюжета и т. д.), чтобы сопоставить их с формальными признаками алгоритма;
2. согласованность оценок: как и в случае с экспертными оценками, можно использовать коэффициент корреляции Пирсона или Спирмена для оценки согласованности между оценками алгоритма и средними оценками читателей. Если читательские оценки выражены в виде категорий (например, "очень плохо", "плохо", "хорошо", "отлично"), можно использовать метрики точности и полноты (precision, recall) для оценки способности алгоритма классифицировать тексты в те же категории, что и читатели;
3. кластерный анализ: на этом шаге следует сгруппировать тексты на основании читательских рейтингов и посмотреть, насколько полученные кластеры совпадают с теми, которые были получены, применив алгоритм. Использование метрики, такой как индекс Джаккарда или силаэтный коэффициент, поможет оценить качество кластеризации и её сходство с читательскими оценками;
4. анализ отзывов читателей: целью этого шага является анализ тональности (sentiment analysis) текстовых отзывов читателей и сопоставить результаты с оценками алгоритма. Например, если читатели используют много позитивных слов, а алгоритм выдает низкую оценку тексту, это может свидетельствовать о необходимости корректировки алгоритма;
5. популярность и доступность: необходимо проверить, насколько оценки алгоритма коррелируют с популярностью текста, которая может быть выражена количеством скачиваний, продаж, или упоминаний в интернете. Высокая корреляция с популярностью может свидетельствовать о том, что алгоритм учитывает характеристики, важные для широкой аудитории;
6. анализ разногласий: оценить тексты, по которым наблюдаются значительные расхождения между читательскими рейтингами и оценками алгоритма. Часто такие расхождения могут выявить важные аспекты восприятия текста, которые не охватываются формальными признаками, например, культурные особенности или эмоциональные реакции, которые могут быть сложны для автоматического анализа;
7. моделирование предпочтений: построить модель на основе читательских данных, например, используя рекомендательные системы. Это может позволить протестировать, насколько формальные признаки, выделенные алгоритмом, связаны с реальными предпочтениями читателей, и могут ли они использоваться для предсказания успешности текста у читательской аудитории.

Необходимо подробнее рассмотреть упоминаемые математические инструменты.

Коэффициент корреляции Пирсона представляет собой количественную меру, применяемую для анализа степени линейной зависимости между двумя переменными. Этот показатель демонстрирует, насколько изменения одной переменной связаны с изменениями другой, а также указывает направление зависимости — прямое или обратное. Этот коэффициент часто обозначается как r и имеет значения в диапазоне от -1 до 1:

r = 1: Полная положительная линейная зависимость. Это значит, что увеличение одной переменной всегда сопровождается пропорциональным увеличением другой;

r=−1: Полная отрицательная линейная зависимость. Увеличение одной переменной сопровождается пропорциональным уменьшением другой;

r=0: Отсутствие линейной зависимости. Это означает, что изменения одной переменной не оказывают влияния на другую (в рамках линейной модели), хотя могут существовать нелинейные зависимости.

Формула для вычисления коэффициента корреляции Пирсона выглядит так:

где:

​ и — значения двух переменных.

и ​ — средние значения переменных x и y.

Коэффициент корреляции Пирсона имеет следующие свойства:

* положительная корреляция: если значение r положительное, это означает, что с увеличением одной переменной увеличивается и другая. Чем ближе r к 1, тем сильнее связь;
* отрицательная корреляция: если значение r отрицательное, это указывает на обратную зависимость, т.е. увеличение одной переменной приводит к уменьшению другой. Чем ближе r к -1, тем сильнее обратная связь.

Отсутствие линейной зависимости: Значение r, близкое к 0, говорит о том, что между переменными нет линейной зависимости. Однако это не исключает возможность нелинейной связи между ними.

Применение коэффициента корреляции Пирсона:

* используется для проверки связи между различными количественными переменными, например, ростом и весом человека;
* применяется для тестирования гипотезы о наличии или отсутствии связи между двумя переменными.

Важно учитывать, что коэффициент Пирсона измеряет линейную зависимость. Это значит, что даже при r=0 между переменными может существовать нелинейная связь, которая не выявляется с помощью данного коэффициента.

Коэффициент корреляции Спирмена представляет собой статистическую меру, применяемую для оценки монотонной зависимости между двумя переменными. Этот непараметрический показатель основывается на анализе ранговых значений, а не исходных числовых данных, что позволяет оценивать степень согласованности переменных в контексте их относительного порядка. В отличие от коэффициента корреляции Пирсона, который измеряет исключительно линейные связи, коэффициент Спирмена способен выявлять зависимости, при которых переменные изменяются согласованно, сохраняя монотонность, даже если связь между ними не является линейной. Благодаря этому он особенно подходит для работы с данными, не подчиняющимися нормальному распределению, или в ситуациях, где данные выражены в виде рангов.

Коэффициент корреляции Спирмена (​) вычисляется по формуле:

где:

​ — разность между рангами значений переменных x и y для каждого наблюдения i.

n — количество наблюдений.

Перед вычислением коэффициента корреляции Спирмена значения каждой переменной преобразуются в ранги, чтобы отразить их относительное положение в наборе данных. Каждому значению присваивается порядковый номер, начиная с 1 для самого низкого значения и заканчивая максимальным значением, которому присваивается наибольший ранг. В случае наличия одинаковых значений в наборе им назначается одинаковый ранг, равный среднему арифметическому рангов для этих позиций.

После ранжирования вычисляется разность ​ между рангами каждой пары наблюдений i для двух переменных. Эта разность используется как основа для последующего расчёта коэффициента корреляции Спирмена, который определяет степень согласованности между ранжированием двух переменных. Этот процесс позволяет выявить монотонные зависимости между переменными, даже если их связь не является линейной.Корреляция рангов: Значение коэффициента изменяется в диапазоне от -1 до 1:

​ = 1 означает полную положительную монотонную зависимость;

= −1 означает полную отрицательную монотонную зависимость;

​ = 0 указывает на отсутствие монотонной зависимости.

Среднеквадратичная ошибка (Mean Squared Error, MSE) является одной из ключевых метрик для оценки точности модели, особенно в задачах регрессии. Она рассчитывается как среднее значение квадратов разностей между предсказанными моделью значениями и фактическими наблюдаемыми данными. MSE помогает определить степень соответствия предсказаний модели реальным данным. Использование квадрата отклонений позволяет усилить влияние крупных ошибок, что делает метрику особенно чувствительной к значительным расхождениям между предсказаниями и наблюдениями. Низкое значение MSE свидетельствует о том, что модель хорошо описывает данные и предоставляет точные прогнозы, минимизируя ошибки. Формула для вычисления MSE:

Где:

— фактическое значение целевой переменной для iii-го наблюдения.

​ — предсказанное значение модели для iii-го наблюдения.

n — количество наблюдений.

Описание:

* фактические и предсказанные значения: MSE измеряет разницу между фактическими значениями () и предсказанными значениями (). Каждое отклонение возводится в квадрат, что делает ошибку всегда положительной и усиливает влияние больших отклонений;
* среднее значение: после возведения в квадрат всех отклонений они суммируются, и результат делится на количество наблюдений (𝑛), чтобы получить среднее значение этих отклонений;
* чувствительность к выбросам: поскольку ошибки возводятся в квадрат, MSE чувствителен к выбросам, так как большие ошибки оказывают большее влияние на итоговое значение.

Средняя абсолютная ошибка (Mean Absolute Error, MAE) представляет собой метрику, применяемую для оценки точности модели. Она рассчитывается как среднее значение абсолютных отклонений предсказанных значений от фактических значений целевой переменной. MAE отражает среднюю величину ошибки модели в тех же единицах измерения, что и данные, и является простым и интуитивно понятным показателем.

Данная метрика широко используется в задачах регрессии благодаря своей интерпретируемости и устойчивости к влиянию крупных ошибок, что отличает её от среднеквадратичной ошибки (MSE). Чем ниже значение MAE, тем точнее модель описывает данные, что свидетельствует о её высокой эффективности в прогнозировании.

Формула для вычисления MAE

Где:

​ — фактическое значение целевой переменной для i-го наблюдения.

​ — предсказанное значение модели для i-го наблюдения.

n — количество наблюдений.

Описание:

* абсолютные отклонения: В отличие от MSE, где отклонения возводятся в квадрат, в MAE используется абсолютное значение разности между фактическими ( ​) и предсказанными (​ ​) значениями. Это делает MAE менее чувствительной к выбросам, так как большие ошибки не усиливаются квадратом.
* среднее значение: Все абсолютные отклонения суммируются, и результат делится на количество наблюдений (n), чтобы получить среднюю абсолютную ошибку. Это дает интуитивное понимание средней ошибки предсказания модели.
* интерпретируемость: MAE легко интерпретировать, так как она показывает среднее отклонение в тех же единицах измерения, что и исходные данные.

Использование этих математических инструментов позволит оценить эффективность полученной оценки.

## Описание основных этапов функционирования многокритериальной оценки художественных текстов

Многокритериальная оценка художественных текстов с использованием лексикографического отношения порядка предусматривает применение нескольких критериев для детального анализа. В данном подходе альтернативы сравниваются последовательно в соответствии с установленным порядком значимости критериев. Сначала рассматривается наиболее важный критерий, и только если альтернативы оказываются равными по этому показателю, анализ продолжается на основе следующего, менее значимого критерия.

Такой метод обеспечивает строгую иерархию приоритетов, что позволяет учитывать различия в значимости критериев при оценке текстов. Это делает лексикографическое отношение порядка особенно эффективным в задачах, где требуется строгое соблюдение заданной последовательности критериев для получения объективных и обоснованных результатов.

Основные этапы функционирования такой оценки:

* определение критериев: на этом этапе определяются критерии, которые будут использоваться для оценки текстов. Это могут быть такие характеристики, как литературное качество, оригинальность сюжета, стиль письма, эмоциональная глубина и т.д.;
* упорядочение критериев: критерии упорядочиваются в соответствии с их важностью. Это делается для того, чтобы определить, какие критерии будут иметь приоритет при принятии решений;
* анализ текстов: каждый текст анализируется с использованием установленных критериев. Это может включать чтение текста, выделение ключевых элементов и оценку соответствия каждого критерия;
* применение лексикографического порядка: после анализа текстов оценки сравниваются между собой с использованием лексикографического порядка. Это означает, что сначала сравниваются значения первого (наиболее важного) критерия. Если значения различаются, то выбирается альтернатива с лучшим значением этого критерия. В случае равенства по первому критерию, сравниваются значения второго критерия, и так далее;
* принятие решения: на основе результатов сравнения текстов принимается решение о том, какой текст лучше всего удовлетворяет установленным критериям. Это может быть выбор одного текста из нескольких альтернатив или ранжирование текстов по степени их соответствия критериям;
* оценка результатов: после принятия решения проводится оценка его соответствия целям оценки и качеству принятого решения. Это позволяет определить эффективность использования лексикографического отношения порядка для оценки текстов;
* адаптация и уточнение: в случае необходимости результаты оценки могут быть использованы для адаптации и уточнения установленных критериев или методов оценки. Это поможет улучшить качество оценки в будущем и обеспечить более точное отражение целей оценки и требований к текстам.

## Описание использования нейронных сетей для многокритериальной оценки художественных текстов

Нейронные сети представляют собой эффективный инструмент для решения задач многокритериальной оценки, где требуется анализировать объекты или ситуации на основе нескольких критериев. Для реализации такого подхода необходимо разработать архитектуру сети, способную обрабатывать входные данные, представляющие значения всех критериев, и генерировать результат, который отражает их совокупное влияние.

Одним из возможных вариантов реализации является создание многовыходной архитектуры нейронной сети. В этом случае каждому критерию соответствует отдельный выход, что позволяет прогнозировать индивидуальные оценки для каждого критерия. Такая сеть обучается на данных, включающих значения всех критериев оценки для каждого объекта или ситуации, обеспечивая детализированный анализ по каждому параметру.

Другой подход основывается на использовании архитектуры с одним выходом. Здесь итоговая оценка рассчитывается как функция от значений всех критериев, поступающих на вход. Для обучения такой сети используется набор данных, содержащий значения критериев и интегральную оценку для каждого объекта. Этот вариант подходит для получения сводной оценки, которая учитывает взаимодействие всех критериев.

Оба подхода обладают своими преимуществами и применяются в зависимости от целей анализа. Многовыходная архитектура лучше подходит для случаев, когда важно рассматривать каждый критерий в отдельности, тогда как одновыходная архитектура фокусируется на генерации интегральной оценки, отражающей общее качество объекта или ситуации.

Основные этапы работы с нейронной сетью:

* работа с данными;
* предварительная обработка;
* конструирование, обучение и оценка качества сети;
* использование и диагностика.

На рисунке 2.1 показана схема обучения нейронной сети, которая может быть использована для анализа данных.

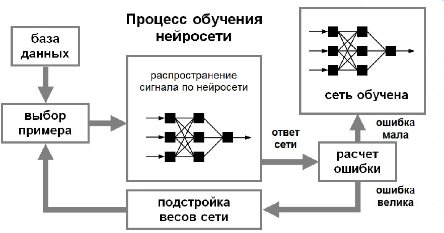


Рисунок 2.1 – Схема обучения нейронной сети

У данного метода есть как преимущества, так и недостатки.

Преимущества:

* расчет зависимости может производиться от нескольких параметров;
* относительно быстрый процесс обучения;
* высокая точность предсказания.

Недостатки:

* принятие решений методом черного ящика.

Задача многокритериального анализа художественного текста требует получения набора независимых, но сопоставимых оценок по различным признакам. Эмоциональность, как один из этих признаков, рассматривается отдельно, с использованием специализированной модели и собственного способа агрегации. При этом подход, реализованный в данной работе, сохраняет независимость оценки эмоциональности от других критериев, что соответствует выбранной архитектуре — построению отдельных моделей для каждого признака. Это позволяет избегать ситуаций, когда ошибка в одном из критериев могла бы повлиять на другие, как это могло бы быть в многозадачной нейросети. Вместе с тем, модель по эмоциональности может быть легко встроена в более крупный пайплайн: её числовой выход (в интервале от 0 до 1) может быть интерпретирован напрямую или позже преобразован в категориальные значения, если потребуется. Важно подчеркнуть, что модель эмоциональности не просто «вычисляет настроение» текста, а оценивает силу выражения эмоций на уровне языка и структуры. Особенность используемого подхода в том, что она основана на анализе предложений, что делает возможным тонкое улавливание эмоционального фона даже в длинных произведениях. За счёт использования предобученной модели, натренированной на русском языке, система может быть масштабирована без необходимости ручной разметки больших корпусов, что критично при работе с художественной литературой.

Таким образом, нейронные сети могут быть использованы для формирования итоговой многокритериальной оценки художественных текстов.

## Выводы по главе 2

В данной главе формализована задача многокритериальная оценка художественных текстов с использованием различных методов, для нахождения оценки каждому критерию будет присвоен вес, после чего будет сформирована финальная оценка.

В рамках исследования определены показатели эффективности, которые позволяют оценивать точность формирования многокритериальной оценки художественных текстов. Эти показатели основаны на сравнении результатов работы модели с экспертными оценками, что обеспечивает объективность и достоверность анализа.

В главе подробно описаны этапы функционирования системы многокритериальной оценки. Рассматривается последовательность действий, начиная с ввода данных и их обработки, до получения итоговой оценки. Уделено внимание ключевым шагам, таким как нормализация данных, выбор критериев и их приоритетов, а также алгоритмы вычислений.

Дополнительно в главе представлено описание процесса обучения нейронных сетей, которые будут использоваться для автоматизации формирования многокритериальной оценки. Рассматриваются подходы к подготовке обучающего набора данных, выбору архитектуры сети и параметров её настройки. Такой подход позволяет обеспечить высокую точность и гибкость системы при оценке художественных текстов с учётом различных критериев.

# ГЛАВА 3. ПРОВЕРКА И ПОДТВЕРЖДЕНИЕ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ ХУДОЖЕСТВЕННЫХ ТЕКСТОВ

## 3.1 Разработка алгоритма анализа художественного текста в виде программного средства.

Python заслуженно считается одним из самых популярных языков программирования для работы с нейронными сетями благодаря его гибкости, интуитивно понятному синтаксису и богатому набору инструментов. Широкий выбор библиотек и фреймворков, таких как TensorFlow, PyTorch, Keras и scikit-learn, значительно упрощает разработку, обучение и внедрение моделей, делая процесс доступным как для новичков, так и для опытных специалистов.

Простота чтения и понимания кода на Python особенно важна при создании сложных проектов или в условиях командной разработки. Обширное сообщество разработчиков активно поддерживает экосистему языка, создавая множество обучающих материалов, руководств и примеров, что облегчает освоение языка и решение возникающих задач. Универсальность Python проявляется и в его совместимости с другими языками, такими как C/C++, Java или R, что позволяет интегрировать Python в различные системы и рабочие процессы.

Для выполнения вычислительно сложных задач Python поддерживает специализированные библиотеки, такие как CUDA и CuDNN, обеспечивающие эффективное использование графических процессоров. Язык также оснащён мощными инструментами для обработки больших объёмов данных, включая Pandas, NumPy и OpenCV, которые помогают с анализом и предобработкой информации. Библиотеки визуализации, такие как Matplotlib и TensorBoard, предоставляют возможности для мониторинга и улучшения процесса обучения моделей. Python идеально подходит для быстрого прототипирования, позволяя проверять идеи и разрабатывать функциональные решения, которые при необходимости можно оптимизировать с использованием других языков. Это делает Python незаменимым инструментом в области разработки нейросетей и анализа данных.

Одним из ключевых преимуществ Python является его платформенная независимость, позволяющая запускать программы на различных операционных системах, включая Windows, macOS и Linux. Это значительно упрощает переносимость приложений и их использование в разнообразных средах. Кроме того, Python активно поддерживается ведущими облачными платформами, такими как AWS, Google Cloud и Azure, что обеспечивает возможность масштабирования вычислительных ресурсов для обучения нейронных сетей и работы с большими объёмами данных.

Благодаря своей универсальности и широким возможностям Python стал ведущим языком в области машинного обучения и искусственного интеллекта. Он предоставляет разработчикам удобный и мощный инструмент, способный решать широкий спектр задач. Одним из ключевых аспектов успеха Python является его доступ к огромному количеству библиотек, которые облегчают реализацию сложных методов и алгоритмов.

В области обработки данных, особенно полученных из Интернета, Python предлагает мощные и универсальные инструменты, такие как библиотеки NumPy и Pandas. Эти библиотеки играют решающую роль в анализе, преобразовании и подготовке данных, необходимых для разработки и обучения моделей, благодаря их высокой производительности и удобству использования.

NumPy (Numerical Python) является одной из ключевых библиотек Python, предназначенной для работы с числовыми данными и решения научных задач. Её основная структура данных, многомерный массив (ndarray), позволяет хранить данные в компактной и однородной форме, обеспечивая при этом высокую производительность за счёт оптимизации на уровне языка C. Благодаря этой архитектуре массивы NumPy значительно превосходят стандартные списки Python по скорости выполнения операций и экономии памяти, что делает их незаменимыми при работе с большими массивами данных.

Библиотека NumPy предоставляет богатый функционал для выполнения разнообразных математических операций, таких как сложение, умножение, статистический анализ и поэлементные вычисления. Помимо этого, она включает инструменты для выполнения линейной алгебры, генерации случайных чисел и преобразований Фурье, которые широко применяются в научных исследованиях и машинном обучении. Её интеграция с такими инструментами, как TensorFlow и PyTorch, а также возможность использования расширений на C/C++ и взаимодействие с языками, такими как Fortran, значительно расширяют сферу её применения.

Кроме того, NumPy является основой для множества других библиотек, включая Pandas и SciPy, которые строят свой функционал на её базовых возможностях. Высокая гибкость и производительность делают NumPy универсальным инструментом, подходящим как для научных вычислений, так и для задач обработки данных и разработки моделей машинного обучения.

Pandas — это универсальная и мощная библиотека Python, предназначенная для анализа данных и работы с табличными структурами, созданная на базе библиотеки NumPy. Она предоставляет удобные инструменты, такие как структуры данных Series и DataFrame, которые делают обработку данных интуитивно понятной и эффективной. Series представляет собой одномерный массив, аналогичный массиву NumPy, но с возможностью использования меток вместо индексов. Это особенно удобно при работе с временными рядами, именованными списками или данными, где важна семантика индексов. DataFrame, в свою очередь, является двумерной структурой, напоминающей таблицы из SQL или электронные таблицы Excel. DataFrame позволяет хранить данные разных типов в виде столбцов, что делает его универсальным инструментом для обработки и анализа. Pandas предоставляет богатый набор функций для работы с данными: загрузка из различных источников (таких как CSV, Excel, JSON и базы данных), фильтрация, сортировка, группировка, агрегация и преобразование. Одной из ключевых особенностей библиотеки является её способность обрабатывать пропущенные значения, что упрощает анализ данных, содержащих пробелы или некорректные записи. Кроме того, Pandas позволяет легко изменять метки строк и столбцов, а также создавать новые столбцы на основе существующих данных. Благодаря своей гибкости и функциональности, Pandas широко используется для предобработки данных, их анализа и подготовки к дальнейшему использованию в моделях машинного обучения или визуализации. Её удобный интерфейс и мощные возможности делают её незаменимым инструментом для работы с данными в самых разных областях.

Pandas прекрасно подходит для решения задач, связанных с предобработкой данных, включая их очистку, объединение различных наборов данных и преобразование форматов. Её инструменты обеспечивают гибкость и эффективность при обработке данных, что делает библиотеку незаменимой на этапе подготовки данных для анализа или моделирования. Библиотека легко интегрируется с другими инструментами Python, такими как Matplotlib для создания визуализаций, SciPy для выполнения сложных математических расчётов и Scikit-learn для задач машинного обучения. Эта совместимость значительно упрощает построение комплексных аналитических решений, объединяющих различные этапы обработки и анализа данных. Одной из сильных сторон Pandas является её функционал для работы с временными рядами. Она поддерживает обработку данных, содержащих даты и временные метки, а также предоставляет инструменты для ресемплирования данных, вычисления скользящих средних и других операций, часто используемых при анализе временных рядов.

Простой и интуитивно понятный синтаксис, в сочетании с богатым функционалом, делает Pandas стандартным инструментом для анализа данных. Она одинаково эффективно работает как с небольшими проектами, так и с крупными массивами данных, включая задачи в распределённых системах. Её универсальность и широкие возможности предопределяют её популярность среди специалистов по обработке данных, аналитиков и разработчиков. Использование описанных выше библиотек позволяет использовать для хранения данных различные форматы данных. Форматы .csv и .xlsx широко используются для хранения и обмена табличными данными, однако их особенности делают их подходящими для разных задач. Формат .csv представляет собой текстовый файл, где данные записаны в виде строк, разделённых запятыми или другим разделителем, например, точкой с запятой. Он прост в использовании, легко читается как человеком, так и программами, но не поддерживает сложное форматирование, стили или встроенные функции. Этот формат подходит для обмена данными между различными системами, так как его можно открыть практически в любом текстовом редакторе или табличном процессоре, например, Microsoft Excel. Однако работа с большими объемами данных или структурированных таблиц вручную в .csv может быть неудобной, поскольку отсутствуют визуальные разделители и форматирование.

Формат .xlsx, как часть экосистемы Microsoft Excel, предоставляет расширенные возможности хранения данных. Он поддерживает сложные таблицы с форматированием, использованием цветов, шрифтов, формул, встроенных объектов и несколькими листами, что делает его удобным для работы человеком, особенно когда требуется анализировать данные в визуально понятной форме. Однако обработка .xlsx с помощью программ требует использования специализированных библиотек, таких как openpyxl или pandas в Python, что снижает его универсальность для автоматизированных процессов по сравнению с .csv. Формат .csv, напротив, благодаря своей простоте и совместимости, более подходит для автоматической обработки данных, в то время как .xlsx оптимален для ситуаций, где приоритетом является анализ и визуализация.

Применение системы контроля версий Git при реализации программного решения для многокритериальной оценки художественных текстов с использованием нейросетей даёт значительные преимущества. Git позволяет эффективно управлять изменениями в коде, архитектурах моделей и данных, что особенно важно в проектах, связанных с машинным обучением. Частые эксперименты с архитектурами нейросетей и параметрами требуют гибкой системы, которая обеспечивает возможность отслеживания каждой внесённой правки. Git также позволяет возвращаться к любой предыдущей версии проекта, что помогает минимизировать риски потери важных наработок и улучшает контроль над ходом разработки. Такой подход делает процесс разработки более прозрачным и устойчивым к ошибкам, способствуя успешной реализации сложных проектов.Кроме того, Git упрощает интеграцию с инструментами для автоматизированного тестирования и развертывания, что критически важно в проекте, включающем нейросети. Регулярная проверка корректности работы кода, качества моделей и производительности системы может быть встроена в процесс разработки, например, через CI/CD. Это гарантирует, что проект остается стабильным даже при активном добавлении новых функций.

Для реализации нейронной сети будет использоваться библиотека TensorFlow.Keras, представляющая собой высокоуровневый интерфейс для разработки моделей глубокого обучения и входящая в состав экосистемы TensorFlow. Основная цель TensorFlow.Keras — облегчить процесс проектирования, обучения и развертывания нейронных сетей. Благодаря удобному и интуитивно понятному API, эта библиотека подходит как для начинающих, так и для опытных специалистов.

Одним из главных преимуществ TensorFlow.Keras является её модульность. Модели создаются из отдельных блоков (слоёв), которые можно легко комбинировать и настраивать. Это позволяет быстро проектировать архитектуры различной сложности, включая многослойные перцептроны, сверточные сети, рекуррентные сети и их комбинации.

Библиотека предоставляет два подхода для определения моделей. Последовательная модель (Sequential) используется для создания простых, линейных архитектур, а функциональный API позволяет реализовывать более сложные модели, включая разветвлённые или объединённые структуры. Такой подход особенно полезен при проектировании моделей с несколькими входами и выходами, а также при создании нетривиальных связей между слоями. TensorFlow.Keras благодаря своей гибкости и широкому набору инструментов существенно упрощает процесс разработки нейронных сетей, делая его доступным и эффективным на всех этапах реализации.

TensorFlow.Keras предоставляет широкий выбор встроенных слоёв, функций активации, оптимизаторов, регуляризаторов и метрик. Это позволяет разработчику сосредоточиться на задаче, не тратя время на реализацию базовых компонентов. Кроме того, Keras упрощает загрузку и предобработку данных, включая работу с популярными наборами данных, такими как MNIST и CIFAR-10, а также поддержку пользовательских форматов данных.

Ещё одной важной особенностью TensorFlow.Keras является поддержка обучения на CPU, GPU и TPU, что делает его подходящим для задач любого масштаба. Также библиотека включает средства для визуализации процесса обучения, например, через TensorBoard, и функции для сохранения и загрузки моделей, что упрощает их переносимость и повторное использование.

TensorFlow.Keras идеально подходит для быстрого прототипирования благодаря удобному интерфейсу и тесной интеграции с остальной частью экосистемы TensorFlow. При этом он достаточно мощный для использования в промышленных приложениях. Благодаря этим характеристикам Keras стал одним из наиболее популярных инструментов для глубокого обучения, предоставляя разработчикам возможность сосредоточиться на решении задач, а не на технических деталях реализации моделей.

Обучение нейронной сети требует предварительной обработки данных. Алгоритм обучения нейронной сети представлен на рисунке 3.1

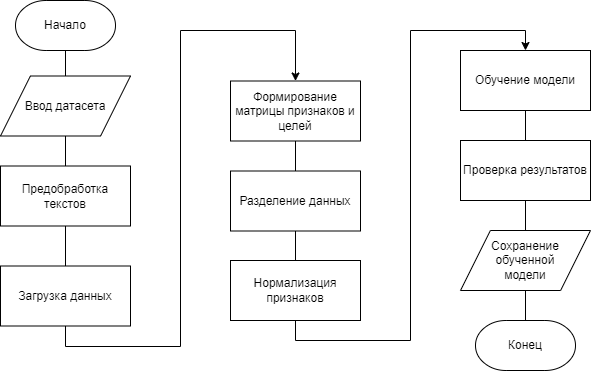


Рисунок 3.1 – Алгоритм обучения нейронной сети

Более подробное описание этапов алгоритма обучения нейронной сети представлено ниже:

1. Ввод датасета

На этом этапе происходит загрузка исходного набора данных, который будет использоваться для обучения нейронной сети (на этом этапе передаются тексты, а также данные об оценках);

1. Предобработка текстов

Если данные содержат текстовую информацию, выполняется её предобработка. Этот этап может включать токенизацию, очистку текста от лишних символов, приведение к нижнему регистру и удаление стоп-слов;

1. Загрузка данных

После подготовки данные загружаются в программу для дальнейшей обработки – на этом этапе для каждого текста вычисляется значения критериев, которые затем будут использоваться в многокритериальной оценке;

1. Формирование матрицы признаков и целей

Из загруженных данных формируется матрица признаков (входных значений) и целей (выходных значений, которые нужно предсказать);

1. Разделение данных

Данные делятся на обучающую и тестовую выборки. Обучающая выборка используется для тренировки модели, а тестовая — для её проверки;

1. Нормализация признаков

Признаки (входные данные) нормализуются для обеспечения корректного обучения модели. Это может включать масштабирование данных или приведение их к определённому диапазону значений;

1. Обучение модели

На этом этапе запускается процесс обучения нейронной сети. Модель использует обучающую выборку, чтобы подстроить свои параметры (веса и смещения);

1. Проверка результатов

После обучения модель тестируется на проверочной выборке, чтобы оценить её точность и качество предсказаний;

1. Сохранение обученной модели

Если результаты модели удовлетворительны, обученная модель сохраняется для дальнейшего использования или развертывания.

Алгоритм работы программы при формировании оценки для нового текста представлен на рисунке 3.2

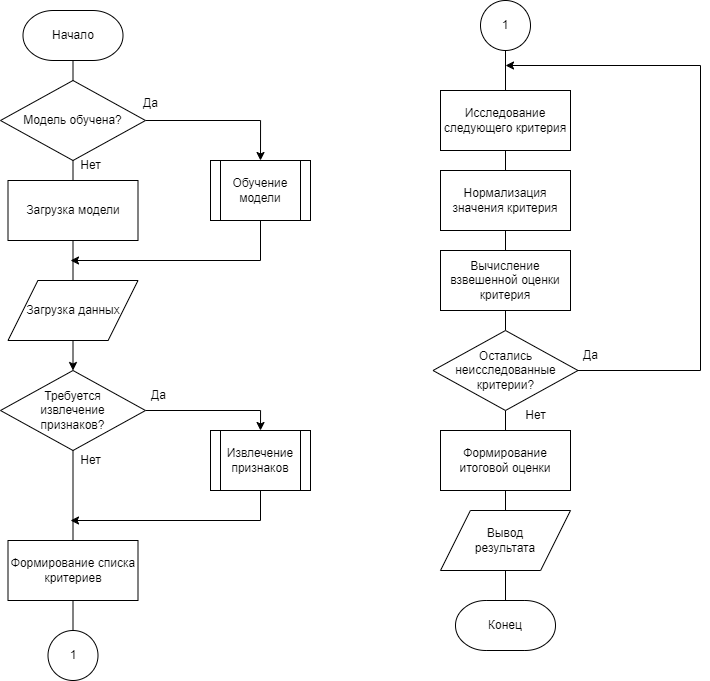


Рисунок 3.2 – Алгоритм работы программы, реализующей многокритериальный анализ художественных текстов с использованием нейросетей

Программа получает на вход текст (или информацию о тексте), производит необходимую обработку и формирует итоговую оценку.

Более подробное описание этапов алгоритма работы программы представлено ниже:

алгоритм условно делится на два основных этапа: предобработка и подготовка данных и многокритериальная оценка:

1. Начало работы

Процесс начинается с проверки текущего состояния модели машинного обучения.

2. Проверка обученности модели

На этом этапе программа определяет, была ли модель уже обучена.Если модель обучена (ответ "Да"), программа переходит сразу к следующему этапу (загрузка данных). Если модель не обучена (ответ "Нет"), выполняется процесс загрузки модели или её обучения.

2.1. Загрузка модели

Если модель уже существует, она загружается в программу.

2.2. Обучение модели

Если модели нет или она не была обучена, запускается этап её обучения на основе загруженных данных.

3. Загрузка данных

После обучения или загрузки модели программа загружает исходные данные, необходимые для анализа и работы с моделью.

4. Проверка необходимости извлечения признаков

На этом этапе программа проверяет, требуется ли выделение дополнительных признаков из загруженных данных. Если извлечение признаков требуется (ответ "Да"), выполняется этап извлечения признаков. Это может включать фильтрацию, преобразование данных, создание новых признаков или проведение анализа на основе исходных данных. Если извлечение признаков не требуется (ответ "Нет"), программа переходит к следующему шагу.

5. Формирование списка критериев

После подготовки данных (с извлечением признаков или без него) формируется список критериев, которые будут использоваться для дальнейшего анализа и оценки. Список критериев задает параметры, по которым программа будет оценивать данные.

6. Исследование критериев (цикл)

На данном этапе программа запускает цикл последовательного исследования каждого критерия из сформированного списка.

6.1. Исследование следующего критерия

Программа анализирует следующий критерий, выбирая его из списка.

6.2. Нормализация значения критерия

Значение критерия приводится к единому масштабу (нормализуется). Нормализация позволяет учесть критерии с разными диапазонами значений и обеспечить их сопоставимость.

6.3. Вычисление взвешенной оценки критерия

Для каждого критерия рассчитывается взвешенная оценка, которая учитывает важность данного критерия в общей оценке. Вес может задаваться заранее или вычисляться программой.

6.4. Проверка наличия не исследованных критериев

Программа проверяет, остались ли еще критерии для анализа: Если есть не исследованные критерии (ответ "Да"), программа возвращается к этапу исследования следующего критерия. Если критериев не осталось (ответ "Нет"), программа переходит к следующему шагу.

7. Формирование итоговой оценки

После завершения анализа всех критериев программа суммирует или агрегирует полученные взвешенные оценки и формирует итоговую оценку. Итоговая оценка отражает общее значение на основе всех рассмотренных критериев.

8. Вывод результата

На данном этапе программа выводит итоговый результат. Это может быть числовая оценка, отчёт или визуализация результатов.

Таким образом, программа получает на вход текст (или информацию о тексте), производит необходимую обработку и формирует итоговую оценку.

В рамках данной работы задача определения эмоциональности художественного текста решается через агрегирование предсказаний нейросетевой модели, обученной на уровне предложений. В основе метода лежит дискретная классификационная модель, предсказывающая одно из пяти значений метки эмоциональности для каждого предложения. Выход модели подвергается переводу в численную шкалу интенсивности, после чего осуществляется агрегирование по всему тексту.

Целевая переменная в задаче оценки эмоциональности представляет собой вещественное значение, лежащее в интервале от 0 до 1. Это значение отражает усреднённую эмоциональную насыщенность текста, определяемую на основе анализа каждого отдельного предложения. В отличие от классической классификации, где эмоциональность приписывается к одному из заранее заданных классов, в данной работе используется непрерывная шкала, что позволяет более точно и гибко оценивать интенсивность эмоционального выражения. Такой подход удобен не только для построения более точных моделей, но и для последующей агрегации результатов по нескольким критериям без жёсткой дискретизации.

Для оценки эмоциональности текста была реализована система, основанная на использовании предварительно обученной модели, представляющей собой реализацию архитектуры BERT, адаптированную под задачу анализа сентимента на русском языке. Модель предсказывает одну из пяти меток эмоциональной окраски для отдельного предложения: от сильно негативной (0) до сильно позитивной (4). Однако в данной работе акцент сделан не на категоризацию каждого высказывания, а на более тонкую и непрерывную оценку интенсивности выражения эмоций во всём тексте.

Процесс анализа начинается с предварительной обработки текста, включающей удаление лишних пробелов, символов переноса строки и разбиение на отдельные предложения. Разметка границ предложений осуществляется с помощью регулярных выражений, ориентированных на знаки препинания (точки, восклицательные и вопросительные знаки). Далее, каждый фрагмент текста проходит через этап лемматизации с использованием библиотеки Natasha, что позволяет привести слова к их начальной форме и очистить текст от лишнего шума, не несущего эмоциональной нагрузки. Эта процедура особенно важна для текстов художественной направленности, где вариативность формы может скрывать повторяющиеся по смыслу конструкции.

Пусть текст состоит из n предложений.

где f() — модель типа AutoModelForSequenceClassification, обученная на задаче сентимент-анализа.

Каждое предложение подаётся на вход модели классификации, которая предсказывает одно из следующих значений: 0 — очень негативный, 1 — негативный, 2 — нейтральный, 3 — позитивный, 4 — очень позитивный. Однако для целей оценки эмоциональной насыщенности учитываются только крайние значения: 0 и 4. Именно они интерпретируются как выражения высокой эмоциональности.

Для интерпретации эмоциональной интенсивности используется весовая функция, согласно которой метки 0 и 4 получают вес 1, а все остальные метки — вес 0. Таким образом, только очень негативные и очень позитивные предложения участвуют в формировании итоговой оценки.

Дополнительно, если предложение заканчивается восклицательным знаком, то его интенсивность считается равной 1 вне зависимости от предсказанной метки. Это обусловлено тем, что восклицательные предложения в художественном тексте, как правило, являются маркерами экспрессии и эмоционального всплеска.

После получения интенсивностей для всех предложений текста, общая оценка эмоциональности вычисляется как среднее значение этих интенсивностей.

Результат находится в диапазоне от 0 до 1, где значение, близкое к нулю, соответствует полностью нейтральному тексту, а значение, близкое к единице, — крайне эмоционально насыщенному произведению.

Стоит отметить, что модель, используемая для классификации предложений, основана на архитектуре BERT и обучена на задаче сентимент-анализа для русского языка.

На уровне внутренних представлений каждое предложение преобразуется в векторное представление, после чего применяется линейный классификатор с функцией активации softmax, возвращающий распределение вероятностей по классам.

Для инференса используется класс с максимальной вероятностью.

Использование среднего значения интенсивности позволяет сохранить чувствительность к распределению эмоций по тексту. Такой подход обеспечивает как гибкость, так и интерпретируемость оценки, особенно при масштабировании анализа на тексты различной длины.

## 3.2 Выбор метрики качества

Поскольку задача сводится к многокритериальной оценке (многомерной регрессии), где для каждого текста мы получаем вектор из нескольких оценок по формальным критериям, ключевым моментом является возможность количественно сравнивать предсказания модели с эталонными (экспертными) значениями.

Основные типы метрик:

* среднеквадратичная ошибка (Mean Squared Error, MSE): применяется к каждому критерию, затем можно усреднить по всем критериям. Итоговая метрика может быть усреднена по всем критериям, также используется RMSE;
* средняя абсолютная ошибка (Mean Absolute Error, MAE);
* коэффициент детерминации (R² Score): можно рассчитать по каждому критерию или в целом (по всем критериям совместно, если иметь дело с векторной формой R²). R² показывает, какую долю вариативности истинных оценок объясняет модель. Чем ближе к 1, тем лучше;
* корреляционные метрики (например, корреляция Пирсона или Спирмана): можно оценить, насколько предсказания ранжируют тексты по каждому критерию схоже с экспертами. Если цель — добиться, чтобы модель воспроизводила не только уровни, но и порядок текстов по критерию, тогда корреляция будет наглядным показателем.

MAE (Mean Absolute Error), или средняя абсолютная ошибка, является одной из наиболее интуитивно понятных и широко используемых метрик в задачах регрессии. Она показывает, насколько в среднем предсказания модели отклоняются от фактических значений целевой переменной. В отличие от других метрик, таких как среднеквадратичная ошибка (MSE) или её корень (RMSE), MAE не усиливает влияние больших ошибок, поскольку использует модуль разности между предсказанием и истиной, а не квадрат. Это делает MAE более устойчивой к выбросам и особенно полезной в тех случаях, когда важно понимать типичную ошибку, совершаемую моделью, без её искусственного увеличения из-за нескольких неудачных прогнозов.

MAE легко интерпретировать: если, например, MAE равна 0.25, это означает, что в среднем модель ошибается на четверть балла (если рейтинг выражен, например, по шкале от 1 до 5). Это особенно важно при работе с оценками художественных текстов, где субъективный фактор играет большую роль, и абсолютная точность модели не всегда достижима. Средняя абсолютная ошибка позволяет оценить практическую применимость модели: если MAE достаточно мала, модель может считаться пригодной для предварительной оценки или фильтрации текстов, даже если она не даёт идеально точных чисел.

Однако MAE имеет и ограничения. Например, в отличие от RMSE, она менее чувствительна к случаям, где ошибка модели существенно выходит за пределы нормы. Поэтому для полноты анализа обычно рекомендуется использовать MAE в сочетании с другими метриками. Формально MAE рассчитывается по следующей формуле (3.3):

|  |  |
| --- | --- |
|  | (3.3) |

RMSE (Root Mean Squared Error) — корень средней квадратичной ошибки. Данная метрика является одной из наиболее распространённых в задачах регрессии и показывает среднюю величину отклонения предсказанных значений от реальных, при этом ошибки возводятся в квадрат, усредняются, а затем извлекается квадратный корень. За счёт квадратичной составляющей RMSE особенно чувствителен к крупным ошибкам модели: чем больше разница между предсказанием и реальностью, тем сильнее это влияет на итоговую метрику. Это делает RMSE полезной для анализа тех моделей, которые в отдельных случаях могут «сильно промахиваться», даже если в среднем работают неплохо.

В задачах анализа художественных текстов данная метрика может помочь выявить модели, которые неустойчивы: они могут неплохо предсказывать для большинства текстов, но ошибаться на тех, где присутствуют сложные синтаксические или стилистические конструкции. Низкое значение RMSE означает, что модель делает стабильные, близкие к реальности предсказания, включая сложные случаи. Способ вычисления этой формулы представлен в формуле 3.4

|  |  |
| --- | --- |
|  | (3.4) |

Оценка качества многокритериальной оценки с помощью метрики RMSE (среднеквадратичная ошибка) заключается в измерении расхождения между результатами, предсказанными моделью, и реальными (эталонными) значениями. Этот подход позволяет понять, насколько точно алгоритм многокритериальной оценки воспроизводит ожидаемые результаты.

R² (коэффициент детерминации)

Коэффициент детерминации R² показывает, какую долю дисперсии зависимой переменной может объяснить модель. Иначе говоря, он измеряет, насколько хорошо предсказанные значения аппроксимируют истинные значения. Если R² равно единице, это означает, что модель идеально объясняет все вариации целевой переменной. Если R² равно нулю, модель не лучше наивного предположения (например, просто среднего значения). Отрицательные значения R² свидетельствуют о том, что модель даёт предсказания хуже, чем простая константа.

Эта метрика особенно полезна при сравнении различных моделей между собой, поскольку она даёт представление о том, какая модель более полно захватывает структуру данных. В задачах анализа художественных текстов она может показать, какая модель лучше интерпретирует сложные, многоуровневые зависимости между признаками (например, между синтаксисом и эмоциональностью текста) и целевой переменной (рейтингом). Вычисление это метрики представлено в формуле 3.5

|  |  |
| --- | --- |
|  | (3.5) |

Процесс оценки качества начинается с подготовки двух наборов данных: предсказанных значений, которые модель или нейронная сеть выдает после обработки входных данных, и истинных значений, представляющих экспертные или эталонные оценки. Каждый из этих наборов содержит числовые оценки для текстов, полученные на основе множества критериев.

Далее для каждой оценки вычисляется разница между предсказанным результатом и эталонным значением. Эти разности возводятся в квадрат, чтобы исключить влияние знаков и усилить влияние больших ошибок. Затем суммируются все квадраты ошибок, и результат делится на количество оценок в выборке. Полученное среднее значение показывает, насколько сильно модель отклоняется от эталонных оценок в среднем.

Чем меньше значение среднеквадратичной ошибки, тем точнее модель многокритериальной оценки. Низкая ошибка говорит о том, что предсказанные значения практически совпадают с истинными, что свидетельствует о высоком качестве модели. Наоборот, если значение ошибки велико, это указывает на значительные расхождения, что требует доработки модели. Причины высокой ошибки могут быть связаны с недостаточным качеством исходных данных, неправильной настройкой параметров модели или слабой архитектурой.

Таким образом, использование MSE для оценки качества многокритериальной оценки позволяет получить числовой показатель точности модели, который легко интерпретировать. Этот подход помогает выявить, насколько эффективно модель справляется с задачей, и является важным инструментом для её дальнейшей оптимизации.

В рамках данной работы для оценки качества разработанных регрессионных алгоритмов использовались три метрики: средняя абсолютная ошибка (MAE), корень средней квадратичной ошибки (RMSE) и коэффициент детерминации (R²). Применение именно этих метрик обусловлено тем, что они позволяют комплексно оценить точность, устойчивость и информативность модели с разных сторон, что особенно важно при анализе художественных текстов, где предсказания носят непрерывный характер и могут варьироваться в широком диапазоне.

Основной метрикой, используемой в процессе анализа, выступает MAE (Mean Absolute Error). Она отражает среднюю величину отклонения предсказанных значений от фактических, измеряемую в тех же единицах, что и целевая переменная. Это делает MAE простой для интерпретации и понятной с точки зрения прикладного анализа: она позволяет судить, насколько сильно модель в среднем ошибается при прогнозировании оценок текстов. Кроме того, MAE устойчива к выбросам, что делает её особенно надёжной при работе с данными, содержащими шум или индивидуальные особенности.

Дополнительно применяется метрика RMSE (Root Mean Squared Error), которая, в отличие от MAE, акцентирует внимание на крупных ошибках модели. Возводя отклонения в квадрат перед усреднением, RMSE увеличивает вес тех случаев, когда предсказание оказалось значительно неточным. Это позволяет зафиксировать нестабильность модели и понять, насколько она склонна к резким отклонениям в предсказаниях даже при хорошем среднем значении ошибки.

Третьей метрикой выбран коэффициент детерминации R², который позволяет определить, какую долю дисперсии целевой переменной можно объяснить на основе построенной модели. Значение R², близкое к единице, указывает на высокую степень соответствия предсказаний реальным значениям, тогда как значение, близкое к нулю или отрицательное, свидетельствует о слабой или отсутствующей зависимости между признаками и целевой переменной. Эта метрика особенно важна для оценки общего качества модели как средства объяснения поведения данных.

Таким образом, использование MAE, RMSE и R² позволяет всесторонне оценить разработанные алгоритмы: с точки зрения средней ошибки, устойчивости к выбросам и объясняющей способности. Такое комбинированное применение метрик обеспечивает более полную и объективную картину эффективности построенных решений.

## Выводы по главе 3

В третьей главе диссертации рассматривается алгоритм автоматической многокритериальной оценки художественных текстов.

Проведен анализ доступных инструментов для реализации алгоритма, и обоснован выбор языка программирования Python и библиотеки машинного обучения Scikit-learn. Представлен алгоритм автоматической кластеризации больших объемов данных, учитывая особенности создания датасетов и обеспечение высокой производительности и точности. Описана программная реализация разработанного алгоритма, включая описание использования библиотек и инструментов программирования. В результате, в третьей главе представлена программная реализация разработанного алгоритма автоматической многокритериальной оценки художественных текстов.

СПИСОК ЛИТЕРАТУРЫ

1. Методические указания по подготовке выпускной квалификационной работы по направлению подготовки бакалавров 09.03.04 "Программная инженерия" / А. А. Доронина, Р. А. Касимов, Е. Л. Федотова; Министерство образования и науки РФ, Национальный исследовательский университет "МИЭТ"; под редакцией Л. Г. Гагариной. - Москва: МИЭТ, 2021. - 28 с.
2. Андриянов, Д. В. Проектирование информационной системы для выборки словарных статей по сти-листическим пометам / Д. В. Андриянов // Актуаль-ные направления научных исследований XXI века:теория и практика. - 2015. -Т. 3, № 7, ч. 3.
3. Zipf G.K. Human Behavior and the Principle of Least Effort. — Addison-Wesley Press, 1949. — С. 484-490. — 573 с.
4. Всеволодова, А. В. Компьютерная обработка лингвистических данных / А. В. Всеволодова. - М. Флинта : Наука, 2007. 96 с
5. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие. М.: Изд-во НИУ ВШЭ,2017. 269 с.