

**САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
ПЕТРА ВЕЛИКОГО**

Физико-механический институт
Высшая школа прикладной математики и вычислительной физики

Реферат на тему

**«Генерация тестовых данных для автоматизированного анализа когнитивных
искажений»
по дисциплине «Автоматизация научных исследований»**

Выполнил: Гулеватая Вероника Вадимовна
студент группы 5040102/50201
Преподаватель: Новиков Ф.А.

Санкт-Петербург
2025 г.

Оглавление

ВВЕДЕНИЕ.....	3
АКТУАЛЬНОСТЬ	3
СТЕПЕНЬ РАЗРАБОТАННОСТИ.....	3
ЦЕЛЬ И ЗАДАЧИ.....	3
ОБЪЕКТ/ПРЕДМЕТ ИССЛЕДОВАНИЯ	4
МЕТОДЫ/ИНСТРУМЕНТЫ	4
ПРАКТИЧЕСКАЯ ЗНАЧИМОСТЬ	5
СТРУКТУРА РАБОТЫ.....	5
ГЛАВА 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ	6
1.1 КОГНИТИВНЫЕ ИСКАЖЕНИЯ: ОПРЕДЕЛЕНИЕ, КЛАССИФИКАЦИЯ, МЕХАНИЗМЫ ВОЗНИКНОВЕНИЯ.....	6
1.2 ЛОЖНАЯ ДИХОТОМИЯ, ИСКАЖЕНИЕ ПОДТВЕРЖДЕНИЯ, ЭФФЕКТ ОРЕОЛА И ЭВРИСТИКА ДОСТУПНОСТИ: ХАРАКТЕРИСТИКИ И ПРОЯВЛЕНИЯ.....	7
1.3 ПОДХОДЫ К АВТОМАТИЗИРОВАННОМУ ВЫЯВЛЕНИЮ КОГНИТИВНЫХ ИСКАЖЕНИЙ	7
1.4 ТРЕБОВАНИЯ К ТЕСТОВЫМ ДАННЫМ ДЛЯ ВАЛИДАЦИИ МЕТОДОВ АНАЛИЗА	9
ГЛАВА 2. ПРОЕКТИРОВАНИЕ СИСТЕМЫ ГЕНЕРАЦИИ ТЕСТОВЫХ ДАННЫХ	11
2.1 ПОСТАНОВКА ЗАДАЧИ ГЕНЕРАЦИИ ТЕСТОВЫХ ДАННЫХ	11
2.2 МОДЕЛИРОВАНИЕ СЦЕНАРИЕВ ПРОЯВЛЕНИЯ КОГНИТИВНЫХ ИСКАЖЕНИЙ.....	12
2.3 СТРАТЕГИИ ГЕНЕРАЦИИ СИНТЕТИЧЕСКИХ ДАННЫХ И ПАРАМЕТРИЗАЦИЯ	14
2.4 СТРУКТУРА ТЕСТОВОГО НАБОРА ДАННЫХ И МЕТАДАННЫЕ	16
ГЛАВА 3. РЕАЛИЗАЦИЯ И ВАЛИДАЦИЯ	18
3.1 ИНСТРУМЕНТЫ И МЕТОДЫ ГЕНЕРАЦИИ ТЕСТОВЫХ ДАННЫХ	18
3.2 ПРИМЕРЫ СГЕНЕРИРОВАННЫХ ДАННЫХ ДЛЯ РАЗЛИЧНЫХ ТИПОВ ИСКАЖЕНИЙ	21
3.3 КРИТЕРИИ КАЧЕСТВА И ВАЛИДАЦИЯ ТЕСТОВЫХ ДАННЫХ	23
3.4 ПРИМЕНЕНИЕ ТЕСТОВЫХ ДАННЫХ ДЛЯ ОЦЕНКИ МЕТОДОВ АВТОМАТИЗИРОВАННОГО АНАЛИЗА	26
ЗАКЛЮЧЕНИЕ	29
СПИСОК ЛИТЕРАТУРЫ	32

Введение

Актуальность

Исследование когнитивных искажений (ложная дихотомия, искажение подтверждения, эффект ореола, эвристика доступности) приобретает особую значимость в контексте автоматизированного анализа человеческого поведения и принятия решений. Разработка надежных методов автоматического выявления этих искажений требует наличия качественных тестовых данных, которые позволяют валидировать алгоритмы и сравнивать различные подходы. Ручное создание таких данных трудоемко, субъективно и не обеспечивает необходимого разнообразия сценариев. Генерация синтетических тестовых данных с контролируемыми параметрами искажений становится практической необходимостью для развития методов автоматизированного анализа когнитивных искажений.

Степень разработанности

Проблематика когнитивных искажений активно изучается в психологии, когнитивной науке и поведенческой экономике. Существуют теоретические модели и эмпирические исследования различных типов искажений. В области автоматизированного анализа развиваются методы обработки естественного языка, извлечения паттернов и классификации текстов. Однако вопросы систематической генерации тестовых данных для валидации методов автоматизированного выявления когнитивных искажений остаются недостаточно разработанными, особенно применительно к специфике различных типов искажений и необходимости обеспечения разнообразия и репрезентативности данных.

Цель и задачи

Цель работы — описать подход к генерации тестовых данных для автоматизированного анализа когнитивных искажений (ложная дихотомия,

искажение подтверждения, эффект ореола, эвристика доступности), включающий моделирование сценариев проявления искажений, генерацию синтетических данных и валидацию качества тестового набора.

Задачи:

- рассмотреть теоретические основы когнитивных искажений и их характеристики;
- определить требования к тестовым данным для автоматизированного анализа;
- описать подходы к моделированию сценариев проявления различных типов искажений;
- представить стратегии генерации синтетических тестовых данных с контролируемыми параметрами;
- описать структуру тестового набора данных и метаданные;
- рассмотреть критерии качества и методы валидации сгенерированных данных;
- продемонстрировать применение тестовых данных для оценки методов автоматизированного анализа.

Объект/предмет исследования

Объект исследования — процесс генерации тестовых данных для автоматизированного анализа когнитивных искажений. Предмет исследования — методы и подходы к моделированию сценариев проявления когнитивных искажений и генерации синтетических тестовых данных с заданными характеристиками.

Методы/инструменты

Методы и инструменты включают анализ теоретических моделей когнитивных искажений, формализацию сценариев проявления искажений, методы генерации синтетических данных (шаблоны, правила, параметризация), структурирование

тестовых наборов с метаданными, методы валидации качества данных (экспертная оценка, статистические метрики, проверка на репрезентативность), применение тестовых данных для оценки алгоритмов автоматизированного анализа.

Практическая значимость

Практическая значимость работы заключается в формировании воспроизводимого подхода к созданию тестовых данных для валидации методов автоматизированного анализа когнитивных искажений. Предложенный подход позволяет систематически генерировать разнообразные тестовые сценарии, контролировать параметры искажений и обеспечивать качество данных для оценки алгоритмов. Результаты могут быть использованы в исследованиях по автоматизированному анализу человеческого поведения, разработке систем поддержки принятия решений и образовательных проектах по когнитивной психологии.

Структура работы

Структура работы включает введение, три главы, заключение и список литературы. В первой главе рассматриваются теоретические основы когнитивных искажений и требования к тестовым данным. Во второй главе описывается проектирование системы генерации тестовых данных, включая моделирование сценариев и стратегии генерации. В третьей главе представлены реализация, примеры сгенерированных данных и валидация качества. В заключении приводятся основные выводы и направления дальнейших исследований.

Глава 1. Теоретические основы

1.1 Когнитивные искажения: определение, классификация, механизмы возникновения

Когнитивные искажения представляют собой систематические отклонения в мышлении, восприятии и принятии решений, возникающие вследствие упрощения процессов обработки информации человеческим мозгом. Термин «когнитивное искажение» (cognitive bias) был введен в научный оборот в работах А. Тверски и Д. Канемана, которые заложили основы поведенческой экономики и когнитивной психологии [1, 2]. Когнитивные искажения не являются случайными ошибками, а представляют собой предсказуемые паттерны мышления, которые возникают в результате эволюционно сформированных эвристик — упрощенных правил принятия решений.

Механизмы возникновения когнитивных искажений связаны с ограниченностью когнитивных ресурсов человека и необходимостью быстрой обработки больших объемов информации. Мозг использует эвристики как «быстрые пути» для принятия решений, что в большинстве ситуаций позволяет эффективно функционировать, но в определенных условиях приводит к систематическим ошибкам. Когнитивные искажения возникают на различных этапах обработки информации: при восприятии, интерпретации, запоминании и принятии решений. Классификация когнитивных искажений может осуществляться по различным основаниям. По характеру влияния на процесс принятия решений выделяют искажения, связанные с обработкой информации (например, искажение подтверждения), с оценкой вероятностей (эвристика доступности), с формированием суждений о людях и объектах (эффект ореола), с категоризацией и выбором (ложная дихотомия). По этапам когнитивного процесса различают искажения восприятия, внимания, памяти и суждения. По области проявления выделяют искажения в социальных суждениях, экономических решениях, оценке рисков и других контекстах.

Важным аспектом изучения когнитивных искажений является их универсальность и устойчивость. Исследования показывают, что когнитивные искажения

проявляются у людей независимо от уровня образования, культурного контекста и профессиональной подготовки, хотя степень выраженности может варьироваться. Это указывает на глубокие эволюционные и нейробиологические корни данных феноменов.

Понимание природы и механизмов когнитивных искажений является фундаментальной основой для разработки методов их автоматизированного выявления. Для создания эффективных тестовых данных необходимо учитывать специфику каждого типа искажения, условия его проявления и характерные паттерны, которые могут быть зафиксированы в текстовых данных, поведенческих траекториях или структурированных ответах.

1.2 Ложная дихотомия, искажение подтверждения, эффект ореола и эвристика доступности: характеристики и проявления

1.3 Подходы к автоматизированному выявлению когнитивных искажений

Автоматизированное выявление когнитивных искажений представляет собой междисциплинарную задачу, находящуюся на стыке обработки естественного языка, психологии, когнитивной науки и машинного обучения. Различные подходы к решению этой задачи могут быть классифицированы по используемым методам и типам анализируемых данных.

Подходы на основе анализа текстовых паттернов используют лингвистические и структурные признаки для выявления характерных маркеров когнитивных искажений. К таким признакам относятся использование определенных лексических конструкций (например, дихотомических формулировок для ложной дихотомии), анализ логической структуры аргументации, выявление асимметрии в представлении информации, обнаружение корреляций между оценками различных аспектов одного объекта. Методы обработки естественного языка, такие как извлечение синтаксических структур, анализ семантических связей, определение тональности и эмоциональной окраски текста, могут быть использованы для автоматического выделения релевантных признаков.

Статистические и машинно-обучающие подходы применяют алгоритмы классификации и регрессии для автоматического выявления когнитивных искажений на основе обучающих данных. Подходы с учителем требуют наличия размеченных данных, где примеры текстов или поведенческих траекторий помечены наличием или отсутствием определенных типов искажений.

Используются различные алгоритмы: от традиционных методов (логистическая регрессия, деревья решений, метод опорных векторов) до современных подходов глубокого обучения (нейронные сети, трансформеры) [6, 7]. Подходы без учителя могут выявлять аномальные паттерны в данных, которые потенциально связаны с когнитивными искажениями.

Подходы на основе анализа структуры аргументации фокусируются на выявлении логических связей между утверждениями, оценкой полноты представления информации и обнаружением пропусков в аргументации. Такие методы могут анализировать наличие контраргументов, баланс представления различных точек зрения, логическую последовательность рассуждений. Для этого используются методы извлечения аргументов, построения графов аргументации и анализа логических структур.

Гибридные подходы комбинируют различные методы для повышения точности и надежности выявления когнитивных искажений. Например, могут сочетаться анализ текстовых паттернов с машинным обучением, статистические методы с экспертными правилами, анализ структуры аргументации с семантическим анализом.

Ключевой проблемой при разработке методов автоматизированного выявления когнитивных искажений является необходимость валидации результатов. Для оценки качества алгоритмов требуются тестовые данные с известными характеристиками искажений, которые позволяют измерять точность, полноту и другие метрики качества. Отсутствие качественных тестовых наборов данных ограничивает развитие методов и затрудняет сравнение различных подходов.

1.4 Требования к тестовым данным для валидации методов анализа

Разработка эффективных методов автоматизированного выявления когнитивных искажений требует наличия качественных тестовых данных, которые позволяют валидировать алгоритмы, сравнивать различные подходы и измерять прогресс в решении задачи. К тестовым данным предъявляются специфические требования, связанные с особенностями когнитивных искажений и задачами их автоматизированного анализа.

Репрезентативность тестовых данных предполагает, что набор примеров должен отражать разнообразие проявлений когнитивных искажений в реальных условиях. Это включает разнообразие контекстов (научные тексты, новостные статьи, социальные сети, переписка), стилей изложения (формальный, неформальный, технический), языковых конструкций и формулировок. Тестовые данные должны включать как явные, так и неявные проявления искажений, различные степени выраженности и комбинации различных типов искажений.

Контролируемость параметров означает возможность точного задания характеристик искажений в тестовых данных. Для каждого примера должны быть известны тип искажения, степень его выраженности, контекст проявления и другие релевантные параметры. Это позволяет не только оценивать общую точность алгоритмов, но и анализировать их работу на различных типах примеров и выявлять систематические ошибки.

Разнообразие и баланс тестового набора данных предполагают наличие достаточного количества примеров для каждого типа искажения, а также примеров без искажений (негативные примеры). Несбалансированные наборы данных могут приводить к смещенным оценкам качества алгоритмов. Важно также включать пограничные случаи, где искажение выражено слабо или проявляется неоднозначно.

Валидность и надежность тестовых данных требуют, чтобы примеры действительно содержали соответствующие типы искажений, а не просто похожие на них паттерны. Валидность может обеспечиваться через экспертную оценку, использование проверенных источников или воспроизведение известных

примеров из литературы. Надежность предполагает согласованность оценок различных экспертов и стабильность характеристик примеров при повторных оценках.

Структурированность и метаданные необходимы для эффективного использования тестовых данных. Каждый пример должен иметь метаданные, включающие тип искажения, контекст, источник, дату создания, версию и другую релевантную информацию. Структурированное представление данных облегчает их использование в различных алгоритмах и обеспечивает воспроизводимость экспериментов.

Масштабируемость тестового набора данных означает возможность его расширения и обновления без потери качества и согласованности. Это особенно важно для долгосрочных исследований и разработки методов, которые должны работать на постоянно растущих объемах данных.

Доступность и воспроизводимость предполагают, что тестовые данные должны быть доступны для исследовательского сообщества и сопровождаться документацией, описывающей методологию их создания, структуру и правила использования. Это обеспечивает возможность сравнения результатов различных исследований и воспроизведения экспериментов.

Удовлетворение перечисленных требований к тестовым данным является сложной задачей, особенно при ручном создании примеров. Генерация синтетических тестовых данных с контролируемыми параметрами представляет собой перспективный подход к решению этой проблемы, позволяющий систематически создавать разнообразные и репрезентативные наборы данных для валидации методов автоматизированного анализа когнитивных искажений.

Глава 2. Проектирование системы генерации тестовых данных

2.1 Постановка задачи генерации тестовых данных

Задача генерации тестовых данных для автоматизированного анализа когнитивных искажений может быть формализована как процесс создания множества примеров, каждый из которых представляет собой текстовый фрагмент или структурированные данные с заданными характеристиками проявления одного или нескольких типов когнитивных искажений. Основная цель генерации заключается в создании репрезентативного, сбалансированного и контролируемого набора данных, который позволяет валидировать алгоритмы автоматизированного выявления искажений и сравнивать различные подходы к решению этой задачи.

Формальная постановка задачи включает следующие компоненты. Входные параметры системы генерации включают: типы когнитивных искажений, для которых необходимо создать примеры (ложная дихотомия, искажение подтверждения, эффект ореола, эвристика доступности); количество примеров для каждого типа искажения; параметры разнообразия (контексты, стили изложения, степени выраженности); требования к балансу набора данных (соотношение примеров с искажениями и без них, распределение по типам искажений); спецификации формата выходных данных и структуры метаданных. Выходные данные системы генерации представляют собой структурированный набор тестовых примеров, где каждый пример содержит: текстовый контент (высказывание, фрагмент текста, диалог или другой формат представления); метаданные с указанием типа искажения, степени выраженности, контекста проявления; дополнительные атрибуты (источник, дата создания, версия, идентификатор); опциональные аннотации и разметку для обучения моделей. Критерии качества генерируемых данных определяются требованиями, рассмотренными в разделе 1.4: репрезентативность, контролируемость параметров, разнообразие и баланс, валидность и надежность, структурированность, масштабируемость, доступность и воспроизводимость.

Система генерации должна обеспечивать возможность контроля и настройки этих параметров.

Ограничения и вызовы задачи генерации тестовых данных связаны с необходимостью балансирования между контролируемостью параметров и естественностью генерируемых примеров. Синтетические данные должны быть достаточно реалистичными, чтобы алгоритмы, обученные на них, могли работать с реальными данными, но при этом иметь четко определенные характеристики искажений для надежной валидации. Дополнительная сложность заключается в необходимости моделирования различных контекстов проявления искажений и обеспечения разнообразия формулировок при сохранении характерных маркеров каждого типа искажения.

Решение задачи генерации тестовых данных требует разработки методологии, которая объединяет формализацию сценариев проявления когнитивных искажений, стратегии генерации синтетических данных и механизмы контроля качества. Такой подход позволяет систематически создавать тестовые наборы данных, удовлетворяющие перечисленным требованиям и пригодные для валидации методов автоматизированного анализа.

2.2 Моделирование сценариев проявления когнитивных искажений

Моделирование сценариев проявления когнитивных искажений является ключевым этапом проектирования системы генерации тестовых данных. Для каждого типа искажения необходимо формализовать типичные ситуации, контексты и паттерны, в которых оно проявляется, что позволяет создавать реалистичные и репрезентативные примеры.

Моделирование ложной дихотомии включает описание ситуаций, в которых сложные проблемы сводятся к выбору между двумя альтернативами. Типичные сценарии включают: принятие решений в условиях неопределенности («либо мы делаем X, либо Y»); оценку людей или явлений в категоричных терминах («хороший или плохой», «правильно или неправильно»); обсуждение политических, социальных или научных вопросов с игнорированием промежуточных позиций. Формализация включает описание дихотомических

конструкций (союзы «либо-либо», «или-или», противопоставления), контекстов применения (деловые решения, личные суждения, публичные дискуссии) и степеней выраженности (явные дихотомии с категоричными формулировками, неявные с подразумеваемым выбором из двух вариантов).

Моделирование искажения подтверждения требует описания ситуаций, в которых информация интерпретируется или отбирается для подтверждения существующих убеждений. Сценарии включают: обсуждение спорных вопросов с односторонней аргументацией; анализ научных или новостных материалов с селективным цитированием; принятие решений на основе подтверждающих фактов при игнорировании противоречащих. Формализация включает описание асимметрии в представлении аргументов (количество и детальность аргументов «за» и «против»), селективности источников, эмоциональной окраски при обсуждении противоречащих фактов, тенденции к укреплению позиции при получении новой информации. Контексты могут включать научные дискуссии, политические дебаты, личные убеждения, профессиональные решения.

Моделирование эффекта ореола предполагает описание ситуаций, в которых общее впечатление влияет на оценку конкретных характеристик. Сценарии включают: оценку людей на основе одного известного аспекта (успех в карьере, внешность, репутация); оценку продуктов или услуг на основе бренда или известности компании; формирование суждений о явлениях на основе первого впечатления. Формализация включает описание корреляций между оценками различных аспектов одного объекта, использование обобщающих характеристик, отсутствие дифференцированной оценки, распространение положительных или отрицательных характеристик на все свойства объекта. Контексты могут включать профессиональные оценки, потребительские решения, социальные суждения, образовательные оценки.

Моделирование эвристики доступности требует описания ситуаций, в которых вероятность или частота оценивается на основе легкости вспоминания примеров. Сценарии включают: оценку рисков на основе недавних или ярких событий; обобщения на основе личного опыта или анекдотических свидетельств;

переоценку значимости медийно освещаемых событий. Формализация включает описание ссылок на недавние события как на типичные примеры, использование анекдотических свидетельств для статистических обобщений, переоценку медийно освещаемых событий, недооценку статистических данных, зависимость оценок от временной близости и эмоциональной значимости событий. Контексты могут включать оценку рисков, принятие решений на основе опыта, анализ трендов, прогнозирование.

Для каждого типа искажения моделирование включает создание библиотеки шаблонов сценариев, описание параметров вариативности (контекст, стиль, степень выраженности) и формализацию правил генерации примеров. Это позволяет систематически создавать разнообразные тестовые данные, которые воспроизводят типичные паттерны проявления когнитивных искажений в различных условиях.

2.3 Стратегии генерации синтетических данных и параметризация

Разработка стратегий генерации синтетических тестовых данных требует выбора подходов, которые обеспечивают баланс между контролируемостью параметров, разнообразием примеров и реалистичностью генерируемого контента.

Рассматриваются несколько стратегий, которые могут применяться отдельно или в комбинации.

Шаблонная генерация основана на использовании predetermined шаблонов с заполняемыми параметрами. Для каждого типа когнитивного искажения создаются шаблоны, отражающие типичные структуры и формулировки.

Например, для ложной дихотомии шаблоны могут включать конструкции типа «Либо [действие А], либо [негативное последствие В]», «[Объект] либо [характеристика 1], либо [характеристика 2]». Параметризация включает подстановку различных значений в слоты шаблона из predetermined словарей (действия, характеристики, объекты, контексты). Преимуществами шаблонной генерации являются высокая контролируемость, возможность точного задания характеристик искажений и воспроизводимость результатов.

Ограничения связаны с потенциальной неестественностью формулировок и ограниченным разнообразием при использовании фиксированных шаблонов. Правила и трансформации предполагают применение правил преобразования к базовым текстам для внесения характеристик когнитивных искажений. Базовые тексты могут быть нейтральными или содержать слабые проявления искажений. Правила трансформации формализуют операции, которые усиливают или добавляют признаки искажений. Например, для искажения подтверждения правила могут включать удаление контраргументов, усиление эмоциональной окраски подтверждающих аргументов, добавление селективных цитат. Для эффекта ореола правила могут добавлять обобщающие характеристики на основе упоминания одного аспекта объекта. Параметризация включает настройку силы применения правил, выбор типов трансформаций и комбинацию различных правил. Этот подход позволяет создавать более естественные примеры, но требует тщательной разработки правил и валидации результатов трансформаций. Генерация на основе моделей языка использует современные языковые модели (например, трансформеры) для создания текстов с заданными характеристиками. Подход включает формулирование промптов, которые направляют модель на генерацию текстов с проявлениями конкретных типов когнитивных искажений. Параметризация включает настройку промптов, температуры генерации, использование контролируемой генерации с ограничениями на структуру или содержание. Преимуществами являются естественность генерируемых текстов и возможность создания разнообразных формулировок. Ограничения связаны с меньшей контролируемостью точных характеристик искажений и необходимостью постобработки и валидации результатов. Гибридные стратегии комбинируют различные подходы для преодоления их ограничений. Например, шаблонная генерация может использоваться для создания структуры, которая затем обогащается с помощью моделей языка. Правила трансформации могут применяться к сгенерированным шаблонами текстам для повышения естественности. Комбинация подходов позволяет использовать преимущества каждого метода и минимизировать их недостатки.

Параметризация системы генерации включает настройку следующих параметров: типы искажений для генерации; количество примеров для каждого типа; распределение по контекстам (научные тексты, новостные статьи, социальные сети, переписка); распределение по стилям изложения (формальный, неформальный, технический); степени выраженности искажений (явные, умеренные, слабые); баланс набора данных (соотношение примеров с искажениями и без них); параметры разнообразия (вариативность формулировок, использование синонимов, различные структуры предложений).

Контроль качества в процессе генерации включает проверку соответствия сгенерированных примеров заданным параметрам, валидацию наличия характерных маркеров искажений, проверку разнообразия и баланса набора данных. Механизмы контроля могут быть автоматическими (на основе правил и метрик) и полуавтоматическими (с привлечением экспертной оценки для выборки примеров).

Выбор стратегии генерации зависит от конкретных требований к тестовым данным, доступных ресурсов и необходимого баланса между контролируемостью и естественностью. В большинстве случаев оптимальным является использование гибридных подходов, которые позволяют систематически создавать разнообразные и репрезентативные тестовые данные с контролируемыми характеристиками.

2.4 Структура тестового набора данных и метаданные

Эффективное использование тестовых данных для валидации методов автоматизированного анализа требует четкой структуры представления данных и наличия подробных метаданных для каждого примера. Структурированное представление облегчает обработку данных алгоритмами, обеспечивает воспроизводимость экспериментов и позволяет анализировать результаты на различных уровнях детализации.

Структура отдельного примера включает следующие компоненты.

Идентификатор — уникальный идентификатор примера, который позволяет однозначно ссылаться на конкретный пример в наборе данных. Текстовый

контент — основной текст примера, который может быть представлен в различных форматах: простое текстовое высказывание, структурированный текст с разметкой, диалог, многоабзачный текст в зависимости от типа искажения и контекста. Метаданные об искажении включают тип когнитивного искажения (ложная дихотомия, искажение подтверждения, эффект ореола, эвристика доступности или комбинация типов), степень выраженности (явная, умеренная, слабая), описание проявления искажения в тексте, указание на конкретные фрагменты текста, демонстрирующие искажение. Контекстные метаданные включают тип контекста (научный текст, новостная статья, социальная сеть, переписка, другое), стиль изложения (формальный, неформальный, технический), тематику (политика, экономика, наука, повседневная жизнь, другое), источник или способ генерации (шаблон, трансформация, модель языка, гибридный). Технические метаданные включают дату создания, версию набора данных, информацию о генераторе (версия, параметры), статус валидации (валидирован экспертом, автоматически проверен, требует проверки).

Структура набора данных включает метаданные на уровне всего набора: общее описание набора данных, версия и дата создания, статистика по типам искажений (количество примеров для каждого типа), распределение по контекстам и стилям, баланс набора (соотношение примеров с искажениями и без них), информация о методологии генерации, требования к использованию и лицензия.

Форматы представления данных могут варьироваться в зависимости от требований использования. JSON-формат обеспечивает структурированное представление с вложенными объектами и массивами, удобен для программной обработки и позволяет хранить сложные метаданные. CSV-формат подходит для простых случаев с плоской структурой метаданных, удобен для анализа в табличных редакторах и статистических пакетах. XML-формат обеспечивает строгую структуру и возможность валидации схемы, подходит для сложных иерархических данных. Специализированные форматы для машинного обучения (например, форматы для популярных библиотек) могут использоваться для прямого применения в алгоритмах обучения.

Аннотации и разметка могут включать дополнительную информацию для обучения моделей: границы фрагментов текста, демонстрирующих искажение; типы искажений на уровне предложений или фраз; эмоциональная окраска и тональность; логическая структура аргументации; связи между различными частями текста. Такая разметка облегчает обучение моделей и позволяет проводить более детальный анализ результатов.

Версионирование и управление изменениями важны для обеспечения воспроизводимости исследований и отслеживания эволюции набора данных.

Каждая версия набора данных должна иметь уникальный идентификатор, описание изменений по сравнению с предыдущей версией, информацию о совместимости с предыдущими версиями и рекомендации по миграции.

Документация набора данных должна включать описание методологии генерации, правила интерпретации метаданных, примеры использования, информацию о валидации и ограничениях набора данных, рекомендации по применению для различных задач. Хорошая документация критически важна для обеспечения правильного использования данных и воспроизводимости результатов исследований.

Правильно структурированный набор тестовых данных с подробными метаданными является основой для эффективной валидации методов автоматизированного анализа когнитивных искажений. Структура должна обеспечивать баланс между детальностью информации и удобством использования, позволяя исследователям легко работать с данными и получать надежные результаты оценки алгоритмов.

Глава 3. Реализация и валидация

3.1 Инструменты и методы генерации тестовых данных

Реализация системы генерации тестовых данных для автоматизированного анализа когнитивных искажений требует выбора и применения соответствующих инструментов и методов, которые обеспечивают эффективное создание

разнообразных и качественных примеров. Рассматриваются различные категории инструментов и их применение в контексте генерации тестовых данных.

Инструменты для шаблонной генерации включают системы управления шаблонами, которые позволяют создавать, хранить и применять шаблоны с параметризацией. Могут использоваться специализированные библиотеки для работы с шаблонами (например, Jinja2 для Python, Mustache для различных языков), системы правил (Drools, CLIPS) или специализированные фреймворки для генерации данных. Ключевые требования к таким инструментам: поддержка параметризации, возможность работы со словарями и базами знаний, генерация вариаций на основе шаблонов, экспорт результатов в требуемые форматы. Для генерации примеров с ложной дихотомией могут использоваться шаблоны с дихотомическими конструкциями, для эффекта ореола — шаблоны с обобщающими характеристиками.

Инструменты для текстовых трансформаций применяются для модификации существующих текстов с целью внесения характеристик когнитивных искажений. Библиотеки обработки естественного языка (NLTK, spaCy для Python, Stanford CoreNLP для Java) позволяют выполнять синтаксический и семантический анализ, на основе которого можно применять правила трансформации. Инструменты для работы с графами зависимостей и синтаксическими деревьями позволяют точно модифицировать структуру предложений. Для искажения подтверждения могут применяться правила удаления контраргументов или усиления подтверждающих аргументов, для эвристики доступности — добавление ссылок на недавние события.

Языковые модели и генеративные системы представляют современный подход к созданию текстов с заданными характеристиками. Могут использоваться предобученные модели (GPT, BERT, T5 и их варианты) через API или локальные инсталляции [8, 9]. Инструменты для контролируемой генерации (например, CTRL, PPLM) позволяют направлять генерацию в нужное русло [10]. Промпт-инжиниринг и настройка параметров генерации (температура, top-k, top-p) используются для контроля разнообразия и качества генерируемых текстов. Для

всех типов искажений могут создаваться специализированные промпты, которые направляют модель на генерацию текстов с соответствующими характеристиками.

Инструменты для управления данными и метаданными включают системы управления базами данных (реляционные, NoSQL), системы версионирования данных (DVC — Data Version Control, Git LFS), инструменты для работы с форматами данных (JSON, XML, CSV парсеры и валидаторы). Важны инструменты для создания и управления метаданными, аннотаций и разметки данных. Системы управления конфигурацией позволяют сохранять параметры генерации для воспроизводимости результатов.

Инструменты контроля качества включают валидаторы на основе правил для проверки соответствия примеров заданным характеристикам, статистические инструменты для анализа распределений и баланса набора данных, инструменты для сравнения с эталонными примерами. Могут использоваться предобученные модели для предварительной оценки наличия искажений в сгенерированных текстах, хотя такие оценки требуют дополнительной валидации.

Интеграционные платформы и пайплайны объединяют различные инструменты в единый процесс генерации. Могут использоваться системы оркестрации (Apache Airflow, Luigi, Prefect) для управления сложными пайплайнами генерации, системы мониторинга для отслеживания качества и производительности, системы логирования для аудита процесса генерации. Контейнеризация (Docker) и виртуализация обеспечивают воспроизводимость окружения и возможность развертывания на различных платформах.

Методология применения инструментов включает этапы: подготовка словарей и баз знаний для параметризации шаблонов; создание библиотеки шаблонов для различных типов искажений; разработка правил трансформации; настройка языковых моделей и промптов; создание пайплайна генерации с интеграцией различных инструментов; реализация механизмов контроля качества; настройка экспорта данных в требуемые форматы; создание системы управления версиями и метаданными.

Выбор конкретных инструментов зависит от требований к масштабу генерации, доступных ресурсов, необходимого уровня контроля и требований к качеству данных. В большинстве случаев оптимальным является комбинирование различных инструментов в гибридном подходе, который использует преимущества каждого метода.

3.2 Примеры сгенерированных данных для различных типов искажений

Для демонстрации возможностей системы генерации тестовых данных и иллюстрации различных типов когнитивных искажений представляются примеры сгенерированных данных с описанием их характеристик и метаданных. Примеры демонстрируют разнообразие контекстов, стилей изложения и степеней выраженности искажений.

Примеры ложной дихотомии демонстрируют различные контексты проявления этого типа искажения. Пример 1 (деловой контекст, явная дихотомия): «Либо мы внедряем новую систему управления немедленно, либо компания потеряет конкурентоспособность в течение года. Третьего варианта не существует — это вопрос выживания бизнеса». Метаданные: тип искажения — ложная дихотомия, степень выраженности — явная, контекст — деловые решения, стиль — формальный. Пример 2 (социальный контекст, умеренная дихотомия): «В этом вопросе можно быть либо на стороне прогресса, либо на стороне устаревших традиций. Промежуточные позиции просто не имеют смысла». Метаданные: тип искажения — ложная дихотомия, степень выраженности — умеренная, контекст — социальные вопросы, стиль — неформальный. Пример 3 (научный контекст, слабая дихотомия): «Исследование показывает, что результаты либо подтверждают гипотезу, либо опровергают её. Других интерпретаций данных не существует». Метаданные: тип искажения — ложная дихотомия, степень выраженности — слабая, контекст — научные дискуссии, стиль — технический. Примеры искажения подтверждения иллюстрируют различные проявления селективного отбора и интерпретации информации. Пример 1 (политический контекст, явное искажение): «Многочисленные исследования подтверждают

эффективность предлагаемой политики. Эксперты единодушно отмечают положительные результаты. Хотя некоторые критики высказывают сомнения, их аргументы основаны на устаревших данных и не заслуживают серьезного внимания». Метаданные: тип искажения — искажение подтверждения, степень выраженности — явная, контекст — политические дискуссии, стиль — формальный, характеристика — асимметрия аргументов, игнорирование контраргументов. Пример 2 (научный контекст, умеренное искажение): «Данные исследования согласуются с нашей гипотезой. Мы видим четкую тенденцию, которая подтверждает первоначальные предположения. Отдельные аномальные значения не влияют на общую картину». Метаданные: тип искажения — искажение подтверждения, степень выраженности — умеренная, контекст — научные исследования, стиль — технический, характеристика — селективная интерпретация данных.

Примеры эффекта ореола демонстрируют влияние общего впечатления на оценку конкретных характеристик. Пример 1 (профессиональный контекст, явный эффект): «Этот руководитель достиг выдающихся результатов в управлении проектами. Очевидно, что он также обладает отличными навыками межличностного общения, глубоким пониманием бизнес-процессов и способностью к стратегическому мышлению. Его подход к решению задач заслуживает самого высокого признания во всех аспектах деятельности».

Метаданные: тип искажения — эффект ореола, степень выраженности — явная, контекст — профессиональные оценки, стиль — формальный, характеристика — распространение положительной оценки на все характеристики. Пример 2 (потребительский контекст, умеренный эффект): «Эта компания известна своими инновационными продуктами. Поэтому логично предположить, что их сервисная поддержка также на высшем уровне, а цены вполне обоснованы». Метаданные: тип искажения — эффект ореола, степень выраженности — умеренная, контекст — потребительские решения, стиль — неформальный.

Примеры эвристики доступности показывают оценку вероятности на основе легкости вспоминания примеров. Пример 1 (оценка рисков, явная эвристика):

«После недавней авиакатастрофы, которая широко освещалась в СМИ, становится очевидным, что полеты на самолетах стали значительно опаснее. Я слышал о нескольких подобных инцидентах в последнее время, что подтверждает рост рисков. Статистика может говорить об обратном, но реальность показывает иное». Метаданные: тип искажения — эвристика доступности, степень выраженности — явная, контекст — оценка рисков, стиль — неформальный, характеристика — переоценка недавних событий, игнорирование статистики. Пример 2 (обобщения на основе опыта, умеренная эвристика): «Судя по моему опыту и рассказам знакомых, большинство людей недовольны своей работой. Это распространенная проблема в современном обществе». Метаданные: тип искажения — эвристика доступности, степень выраженности — умеренная, контекст — социальные обобщения, стиль — неформальный, характеристика — обобщение на основе анекдотических свидетельств.

Примеры комбинаций искажений демонстрируют случаи, когда в одном тексте проявляются несколько типов искажений. Пример: «Либо мы поддерживаем эту политику, либо отказываемся от прогресса (ложная дихотомия). Все исследования подтверждают её эффективность, а критики просто не понимают сути вопроса (искажение подтверждения). Лидеры, которые её продвигают, заслуживают доверия во всех аспектах (эффект ореола)». Метаданные: типы искажений — ложная дихотомия, искажение подтверждения, эффект ореола, контекст — политические дискуссии.

Представленные примеры иллюстрируют разнообразие проявлений когнитивных искажений и демонстрируют возможности системы генерации по созданию репрезентативных тестовых данных для различных контекстов и степеней выраженности.

3.3 Критерии качества и валидация тестовых данных

Обеспечение качества сгенерированных тестовых данных является критически важным аспектом системы генерации, так как от качества данных напрямую зависит надежность валидации методов автоматизированного анализа.

Разрабатываются критерии качества и процедуры валидации, которые позволяют оценивать и контролировать характеристики генерируемых данных.

Критерии качества на уровне отдельного примера включают следующие аспекты.

Корректность типа искажения означает, что пример действительно демонстрирует заявленный тип когнитивного искажения, а не просто похожие паттерны.

Проверка включает анализ наличия характерных маркеров искажения, соответствия определению типа искажения, отсутствия противоречий с метаданными. Соответствие степени выраженности требует, чтобы пример демонстрировал заявленный уровень выраженности искажения (явный, умеренный, слабый). Для явных искажений должны присутствовать четкие и однозначные признаки, для слабых — более тонкие проявления. Естественность формулировок предполагает, что текст выглядит естественно и не содержит явных признаков искусственной генерации. Проверка включает оценку грамматической корректности, стилистической согласованности, логической связности текста. Релевантность контекста означает соответствие примера заявленному контексту (научный, деловой, социальный и т.д.) и стилю изложения (формальный, неформальный, технический).

Критерии качества на уровне набора данных включают требования к репрезентативности, балансу и разнообразию. Репрезентативность оценивается через разнообразие контекстов, стилей, формулировок и степеней выраженности в наборе данных. Набор должен отражать реальное разнообразие проявлений когнитивных искажений. Баланс набора данных требует наличия достаточного количества примеров для каждого типа искажения, а также примеров без искажений (негативные примеры). Несбалансированные наборы могут приводить к смещенным оценкам алгоритмов. Разнообразие оценивается через метрики лексического разнообразия, вариативности структур предложений, различий в формулировках при сохранении характерных маркеров искажений. Полнота покрытия означает наличие примеров для всех требуемых комбинаций параметров (типы искажений × контексты × стили × степени выраженности).

Методы валидации включают автоматические и экспертные процедуры.

Автоматическая валидация использует правила и алгоритмы для проверки соответствия примеров заданным критериям. Могут применяться проверки на наличие характерных маркеров искажений (ключевые слова, структуры предложений), грамматическая проверка, проверка метаданных на полноту и корректность, статистический анализ распределений и баланса набора данных. Автоматическая валидация позволяет быстро обрабатывать большие объемы данных, но может пропускать тонкие нюансы.

Экспертная валидация предполагает оценку примеров специалистами в области когнитивной психологии, обработки естественного языка или предметной области. Эксперты оценивают корректность типа искажения, степень выраженности, естественность формулировок, релевантность контекста. Экспертная валидация может быть полной (оценка всех примеров) или выборочной (оценка репрезентативной выборки). Для обеспечения надежности используется оценка несколькими экспертами с измерением межэкспертной согласованности (например, коэффициент Каппа Коэна) [11].

Гибридные методы валидации комбинируют автоматические и экспертные процедуры. Автоматическая валидация используется для первичной фильтрации и выявления явных проблем, эксперты оценивают пограничные случаи и сложные примеры. Предобученные модели могут использоваться для предварительной оценки, но результаты требуют экспертной проверки.

Метрики качества позволяют количественно оценивать характеристики набора данных. Метрики на уровне примера включают оценки корректности, естественности, релевантности. Метрики на уровне набора включают распределение по типам искажений, контекстам, стилям; лексическое разнообразие; баланс классов; покрытие параметрического пространства. Статистические метрики (средние, дисперсии, распределения) позволяют оценивать репрезентативность и разнообразие.

Процедуры контроля качества включают этапы: автоматическая проверка при генерации; первичная фильтрация некорректных примеров; экспертная оценка

выборки; статистический анализ набора данных; итеративное улучшение на основе результатов валидации; документирование процедур валидации и результатов.

Управление качеством предполагает создание системы отслеживания качества на всех этапах генерации, ведение логов валидации, управление версиями наборов данных с указанием результатов валидации, создание отчетов о качестве для пользователей данных.

Эффективная система валидации обеспечивает создание надежных тестовых данных, которые могут быть использованы для валидации методов автоматизированного анализа когнитивных искажений с уверенностью в их качестве и репрезентативности.

3.4 Применение тестовых данных для оценки методов автоматизированного анализа

Сгенерированные тестовые данные предназначены для валидации и оценки методов автоматизированного выявления когнитивных искажений.

Рассматриваются подходы к применению тестовых данных для различных задач оценки и анализа эффективности алгоритмов.

Оценка точности классификации является основной задачей применения тестовых данных. Тестовый набор с известными метками типов искажений используется для оценки способности алгоритма правильно идентифицировать наличие и тип когнитивного искажения в тексте. Метрики оценки включают точность (accuracy), полноту (recall), точность (precision), F-меру для каждого типа искажения и для набора в целом. Матрица ошибок (confusion matrix) позволяет анализировать систематические ошибки алгоритма, выявлять типы искажений, которые часто путаются между собой, и определять направления улучшения.

Анализ работы на различных подмножествах данных позволяет оценить, как алгоритм работает в различных условиях. Анализ по контекстам показывает эффективность на научных текстах, новостных статьях, социальных сетях и других типах контекстов. Анализ по стилям изложения выявляет зависимость качества от формальности или неформальности текста. Анализ по степеням

выраженности показывает, насколько хорошо алгоритм справляется с явными, умеренными и слабыми проявлениями искажений. Такой анализ позволяет выявить ограничения алгоритма и области, требующие улучшения.

Оценка устойчивости и обобщающей способности проверяет, насколько хорошо алгоритм работает на данных, которые отличаются от обучающих. Тестовые данные с контролируемыми параметрами позволяют систематически варьировать характеристики примеров и оценивать, как изменения влияют на работу алгоритма. Анализ устойчивости включает проверку работы на различных формулировках, синонимах, вариациях структур предложений при сохранении характерных маркеров искажений.

Сравнение различных подходов является важным применением тестовых данных. Единый тестовый набор позволяет объективно сравнивать различные алгоритмы и методы выявления когнитивных искажений. Сравнение может включать различные типы моделей (правила, статистические методы, глубокое обучение), различные архитектуры нейронных сетей, различные подходы к извлечению признаков. Единый набор данных обеспечивает справедливое сравнение и выявление преимуществ и недостатков каждого подхода.

Анализ ошибок и улучшение алгоритмов использует тестовые данные для детального изучения случаев, когда алгоритм работает неправильно. Анализ ложных срабатываний (false positives) и пропусков (false negatives) позволяет выявить систематические паттерны ошибок, понять причины неправильной классификации и разработать улучшения. Примеры из тестового набора с известными характеристиками позволяют точно определить, какие типы примеров вызывают проблемы.

Оценка работы на комбинациях искажений проверяет способность алгоритма выявлять случаи, когда в тексте присутствуют несколько типов искажений одновременно. Тестовые данные с комбинациями искажений позволяют оценить, может ли алгоритм идентифицировать все присутствующие типы или склонен фокусироваться на одном типе.

Калибровка и настройка параметров алгоритмов использует тестовые данные для оптимизации гиперпараметров моделей, пороговых значений для классификации, весов различных признаков. Разделение тестового набора на обучающую, валидационную и тестовую части позволяет проводить систематическую настройку без переобучения.

Бенчмаркинг и отслеживание прогресса использует стандартизированные тестовые наборы для измерения прогресса в разработке методов автоматизированного анализа. Публично доступные тестовые наборы с известными характеристиками позволяют исследователям сравнивать свои результаты с предыдущими работами и отслеживать улучшения в области. Образовательные и демонстрационные цели включают использование тестовых данных для обучения студентов и исследователей, демонстрации работы алгоритмов, иллюстрации различных типов когнитивных искажений. Примеры с известными характеристиками позволяют наглядно показать, как алгоритм идентифицирует искажения и какие признаки он использует.

Эффективное применение тестовых данных для оценки методов автоматизированного анализа требует тщательного планирования экспериментов, правильного использования метрик оценки, систематического анализа результатов и документирования процедур и выводов. Качественные тестовые данные являются основой для надежной валидации алгоритмов и прогресса в области автоматизированного выявления когнитивных искажений.

Заключение

В настоящей работе рассмотрен подход к генерации тестовых данных для автоматизированного анализа когнитивных искажений, включающий теоретические основы, проектирование системы генерации и методы валидации. Проведенное исследование позволяет сформулировать следующие основные выводы.

Во-первых, систематическая генерация тестовых данных с контролируемыми параметрами представляет собой эффективный подход к решению проблемы недостатка качественных данных для валидации методов автоматизированного выявления когнитивных искажений. В отличие от ручного создания примеров, автоматизированная генерация позволяет создавать разнообразные, сбалансированные и репрезентативные наборы данных с заданными характеристиками.

Во-вторых, успешная генерация тестовых данных требует глубокого понимания природы и механизмов проявления различных типов когнитивных искажений. Для каждого типа искажения (ложная дихотомия, искажение подтверждения, эффект ореола, эвристика доступности) необходимо формализовать типичные сценарии проявления, контексты применения и характерные лингвистические маркеры, что позволяет создавать реалистичные и репрезентативные примеры.

В-третьих, гибридные стратегии генерации, комбинирующие шаблонную генерацию, правила трансформации и использование языковых моделей, обеспечивают оптимальный баланс между контролируемостью параметров, разнообразием примеров и естественностью генерируемых текстов. Такой подход позволяет систематически создавать тестовые данные, удовлетворяющие требованиям репрезентативности, валидности и масштабируемости.

В-четвертых, эффективная валидация сгенерированных данных требует комбинации автоматических и экспертных процедур. Автоматическая валидация обеспечивает быструю обработку больших объемов данных и проверку соответствия формальным критериям, в то время как экспертная оценка

необходима для проверки корректности типа искажения, естественности формулировок и релевантности контекста.

В-пятых, правильно структурированные тестовые данные с подробными метаданными являются основой для надежной валидации методов автоматизированного анализа. Структурированное представление данных, включающее метаданные об искажениях, контекстах, стилях и способах генерации, облегчает использование данных в различных алгоритмах и обеспечивает воспроизводимость экспериментов.

Практическая значимость работы заключается в формировании воспроизводимого подхода к созданию тестовых данных, который может быть использован в исследованиях по автоматизированному анализу человеческого поведения, разработке систем поддержки принятия решений и образовательных проектах. Предложенная методология позволяет систематически генерировать разнообразные тестовые сценарии, контролировать параметры искажений и обеспечивать качество данных для оценки алгоритмов.

Ограничения работы связаны с необходимостью дальнейшей разработки и валидации конкретных инструментов и методов генерации, расширения библиотек шаблонов и правил трансформации, а также накопления опыта применения сгенерированных данных для валидации различных алгоритмов.

Дополнительные исследования могут быть направлены на разработку специализированных языковых моделей для генерации текстов с когнитивными искажениями, создание автоматизированных систем валидации на основе машинного обучения и расширение набора типов искажений, для которых генерируются тестовые данные.

Направления дальнейших исследований включают: разработку методов генерации данных для других типов когнитивных искажений; создание адаптивных систем генерации, которые могут автоматически настраивать параметры на основе результатов валидации алгоритмов; исследование эффективности различных стратегий генерации для различных типов искажений и контекстов; разработку стандартов и протоколов для обмена тестовыми данными между

исследователями; создание публично доступных наборов тестовых данных для бенчмаркинга методов автоматизированного анализа.

В целом, предложенный подход к генерации тестовых данных для автоматизированного анализа когнитивных искажений представляет собой перспективное направление исследований, которое может способствовать развитию методов автоматизированного выявления когнитивных искажений и повышению качества анализа человеческого поведения и принятия решений.

Список литературы

1. Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases / A. Tversky, D. Kahneman // Science. — 1974. — Vol. 185, № 4157. — P. 1124—1131.
2. Kahneman, D. Thinking, fast and slow / D. Kahneman. — New York : Farrar, Straus and Giroux, 2011. — 499 p.
3. Thorndike, E. L. A constant error in psychological ratings / E. L. Thorndike // Journal of Applied Psychology. — 1920. — Vol. 4, № 1. — P. 25—29.
4. Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises / R. S. Nickerson // Review of General Psychology. — 1998. — Vol. 2, № 2. — P. 175—220.
5. Oswald, M. E. Confirmation bias in cognitive psychology / M. E. Oswald, S. Grosjean // Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory / ed. by R. F. Pohl. — Hove : Psychology Press, 2004. — P. 79—96.
6. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M. Chang, K. Lee, K. Toutanova // Proceedings of NAACL-HLT. — 2019. — Vol. 1. — P. 4171—4186.
7. Vaswani, A. Attention is all you need / A. Vaswani [et al.] // Advances in Neural Information Processing Systems. — 2017. — Vol. 30. — P. 5998—6008.
8. Brown, T. Language models are few-shot learners / T. Brown [et al.] // Advances in Neural Information Processing Systems. — 2020. — Vol. 33. — P. 1877—1901.
9. Raffel, C. Exploring the limits of transfer learning with a unified text-to-text transformer / C. Raffel [et al.] // Journal of Machine Learning Research. — 2020. — Vol. 21. — P. 1—67.
10. Keskar, N. S. CTRL: A Conditional Transformer Language Model for Controllable Generation / N. S. Keskar [et al.] // arXiv preprint arXiv:1909.05858. — 2019.
11. Cohen, J. A coefficient of agreement for nominal scales / J. Cohen // Educational and Psychological Measurement. — 1960. — Vol. 20, № 1. — P. 37—46.
12. Hastie, T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction / T. Hastie, R. Tibshirani, J. Friedman. — 2nd ed. — New York : Springer, 2009. — 745 p.
13. Manning, C. D. Foundations of Statistical Natural Language Processing / C. D. Manning, H. Schütze. — Cambridge : MIT Press, 1999. — 680 p.
14. Bird, S. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit / S. Bird, E. Klein, E. Loper. — Sebastopol : O'Reilly Media, 2009. — 504 p.
15. Gilovich, T. Heuristics and biases: The psychology of intuitive judgment / T. Gilovich, D. Griffin, D. Kahneman. — Cambridge : Cambridge University Press, 2002. — 857 p.