Глава 1. Теоретические и практические основы детекции ИИ-генерированного контента

Цель главы — обосновать научную и социальную необходимость разработки методологии, определить ключевые проблемы и существующие подходы, а также сформулировать концептуальную основу проекта.

1.1. Актуальность проблемы распространения ИИ-генерированного контента

Рассматривается рост объема контента, созданного языковыми моделями, и его влияние на информационную среду, включая манипуляции, дезинформацию и подрыв доверия к медиа.

1.1.1. Статистика распространения фейков и синтетического контента в СМИ и соцсетях

Приводятся данные исследований о росте числа фейковых новостей, поддельных отзывов и автоматизированных постов, созданных ИИ, с акцентом на русскоязычное пространство.

1.1.2. Социальные и экономические последствия использования ИИ-контента

Анализируются риски для общественного доверия, политической стабильности, репутации брендов и институтов, а также финансовые потери от манипуляций.

1.2. Обзор существующих решений и их ограничения

Анализируются современные инструменты и алгоритмы детекции ИИ-текста, выявляются их слабые стороны и пробелы, особенно в контексте русского языка.

1.2.1. Обзор зарубежных детекторов и их эффективность

Оценивается точность, доступность и адаптация под английский язык, а также отсутствие поддержки русскоязычного контента.

1.2.2. Российские аналоги и их особенности (например, детекторы от Сбера, Яндекса, НТО)

Анализируются закрытые и частично доступные решения, их ограниченная прозрачность и отсутствие открытых методологий.

1.2.3. Критика существующих подходов: переобучение, стилистическая адаптация, уязвимость к рерайтингу

Выявляются системные недостатки: снижение точности при редактировании текста, зависимость от конкретных моделей, отсутствие объяснимости решений.

1.3. Концептуальная модель разрабатываемой системы

Формулируется авторский подход к детекции, основанный на лингвистических признаках с акцентом на интерпретируемость и адаптацию к русскому языку.

1.3.1. Принципы лингвистической детекции: стилистика, синтаксис, семантическая когерентность

Описывается, как анализ структуры предложений, повторов, логических связей и лексического разнообразия позволяет выявлять “машинность” текста.

1.3.2. Архитектурная схема системы: от входных данных до вердикта

Представляется общая схема обработки: препроцессинг → извлечение признаков → классификация → генерация объяснения.

1.3.3. Отличия предлагаемой методологии от существующих подходов

Подчеркивается открытость, ориентация на русский язык, объяснимость и независимость от конкретных генеративных моделей.

Глава 2. Разработка и реализация методологии детекции ИИ-контента

Цель главы — представить техническую реализацию методологии: от сбора данных и построения модели до создания прототипа веб-интерфейса.

2.1. Формирование корпуса данных для обучения и тестирования

Описывается процесс сбора, разметки и нормализации выборок реальных и сгенерированных текстов на русском языке.

2.1.1. Источники реальных текстов: литература, СМИ, блоги, научные статьи

Перечисляются открытые источники (Проект «Гутенберг», «Литрес», «Медуза», «Хабр»), критерии отбора и очистки данных.

2.1.2. Генерация синтетических текстов с помощью LLM: GigaChat, YaGPT, RuGPT

Описываются параметры генерации (длина, стиль, тематика), чтобы обеспечить разнообразие и реалистичность синтетических данных.

2.1.3. Разметка и валидация выборок: экспертная оценка, автоматическая фильтрация, баланс классов

Рассказывается о применении экспертных оценок и автоматических метрик (например, perplexity) для повышения качества разметки.

2.2. Разработка лингвистической модели классификации

Детально описывается процесс построения классификатора на основе лингвистических признаков, выбор алгоритмов и метрик оценки.

2.2.1. Извлечение лингвистических признаков: частотность конструкций, энтропия, длина предложений, лексическое разнообразие

Перечисляются и обосновываются ключевые признаки, используемые для дифференциации “человеческого” и “машинного” стиля.

2.2.2. Выбор и обучение классификатора: сравнение логистической регрессии, SVM, градиентного бустинга и нейросетей

Приводятся результаты экспериментов по точности, скорости и интерпретируемости различных моделей на тестовом корпусе.

2.2.3. Метрики оценки эффективности: precision, recall, F1-score, AUC-ROC, калибровка вероятностей

Обосновывается выбор метрик, учитывающих как общую точность, так и способность системы давать надежные вероятностные оценки.

2.3. Реализация веб-интерфейса и сценарии взаимодействия

Описывается архитектура и функциональность пользовательского интерфейса, а также основные сценарии использования системы.

2.3.1. Технологический стек

Указываются технологии фронтенда, бэкенда и развертывания, обеспечивающие масштабируемость и удобство использования.

2.3.2. Основной сценарий: загрузка текста → анализ → вывод вероятности + объяснение признаков

Детально описывается пользовательский путь: ввод текста, получение вердикта (0.0–1.0) и интерпретации (“высокая частота пассивных конструкций”, “низкое лексическое разнообразие” и т.п.).

2.3.3. Дополнительные функции: история проверок, API для интеграции, экспорт отчетов

Рассматриваются расширенные возможности: сохранение истории, интеграция с CMS или редакторами, генерация PDF-отчетов с анализом.

Глава 3. Практическая значимость, внедрение и перспективы развития

Цель главы — продемонстрировать ценность проекта для заинтересованных сторон, предложить пути внедрения и обозначить направления дальнейших исследований.

3.1. Практическая польза для целевых групп пользователей

Определяются конкретные выгоды для журналистов, редакторов, модераторов, образовательных учреждений и рядовых пользователей.

3.1.1. Для СМИ и журналистов: верификация источников, защита от плагиата и фейков

Система помогает быстро проверять материалы на “искусственность”, снижая риски публикации недостоверного контента.

3.1.2. Для образовательных учреждений: выявление ИИ-генерированных работ студентов

Позволяет преподавателям объективно оценивать оригинальность письменных работ, сохраняя академическую честность.

3.2. Возможности интеграции и масштабирования решения

Рассматриваются пути внедрения системы в существующие платформы и инфраструктуры, а также адаптация под другие языки и форматы.

3.2.1. Адаптация под другие языки и типы контента (аудио, видео субтитры)

Обсуждаются перспективы расширения методологии на английский, украинский языки, а также на транскрипты и субтитры.

3.2.2. Открытая публикация методологии и модели для научного использования

Подчеркивается ценность открытого доступа к алгоритмам и весам модели для развития сообщества и создания производных решений.

3.3. Ограничения и направления дальнейших исследований

Анализируются текущие слабые места системы и формулируются задачи для будущих улучшений.

3.3.1. Зависимость от качества выборок: риск переобучения, смещение в сторону конкретных генераторов

Обсуждается необходимость постоянного обновления обучающего набора и включения новых моделей ИИ.

3.3.2. Уязвимость к adversarial-атакам и намеренному “очеловечиванию” текста

Рассматриваются методы обхода детектора и пути повышения устойчивости.

3.3.3. Перспективы: комбинированные модели (лингвистика + статистика + поведенческие признаки)

Формулируется гипотеза о повышении точности за счет объединения лингвистических признаков с анализом скорости набора, метаданных и стилистических шаблонов.