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

1.1. Актуальность проблемы распространения ИИ-генерированного контента

1.1.1. Статистика распространения фейков и синтетического контента в СМИ и соц. сетях

1.1.2. Социальные и экономические последствия использования ИИ-контента

1.2. Обзор существующих решений и их ограничения

1.2.1. Обзор зарубежных детекторов и их эффективность

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

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

1.3. Концептуальная модель разрабатываемой системы

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

1.3.2. Архитектурная схема системы: от входных данных до вердикта

1.3.3. Отличия предлагаемой методологии от существующих подходов

Глава 2. Разработка и реализация методологии детекции ИИ-контента

2.1. Формирование корпуса данных для обучения и тестирования

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

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

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

2.2. Разработка лингвистической модели классификации

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

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

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

2.3. Реализация веб-интерфейса и сценарии взаимодействия

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

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

Глава 3. Практическая значимость, внедрение и перспективы развития

3.1. Практическая польза для целевых групп пользователей

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

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

3.2. Возможности интеграции и масштабирования решения

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

3.2.2. Открытая публикация методологии и модели для научного использования

3.3. Ограничения и направления дальнейших исследований

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

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

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