ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

Отчет о командном программном проекте на тему:

Анализ и классификация текстов с целью выявления авторства искусственного интеллекта: исследование лексических, синтаксических и стилистических особенностей текстов, сгенерированных нейросетями

(промежуточный, этап 1)

Выполнили студенты:

группы #БПМИ236, 2 курса	Синицина Софья Андреевна
группы #БПМИ236, 2 курса	Совкова София Денисовна

Принял руководитель проекта:

Галицкий Борис Васильевич
Приглашенный преподаватель
Факультет компьютерных наук НИУ ВШЭ

Москва 2025

Содержание

Аннотация

Проект Detect-AI-Generated-Text направлен на разработку системы для автоматического определения, был ли текст сгенерирован искусственным интеллектом. С использованием нейросетевой модели, построенной на TensorFlow и Keras, система анализирует текстовые данные и предсказывает вероятность их принадлежности к категории AI-генерированных.

Задача проекта заключается в разработке модели, которая эффективно классифицирует тексты как созданные ИИ или человеком. Для этого используется метод TF-IDF для векторизации текста и нейросеть с полносвязными слоями для предсказания вероятности.

К текущему моменту достигнуты следущие результата: созданая и обученая модель на основе двух датасетов: train-essays.csv (данные с Kaggle) и train-v2-drcat-02.csv (дополнительные тексты). В дальнейшем датасеты могут быть использованы другие. Модель продемонстрировала точность 99% на открытых тестах и 82% на соревновании Kaggle. В будущем алгоритм будет интегрирован в Telegram-бота для анализа текстов в реальном времени, что позволит пользователям получать прогноз вероятности искусственного происхождения текста.

Ключевыми особенностями подхода являются использование TF-IDF для векторизации текста и нейросетевой архитектуры с тремя полносвязными слоями и активацией ReLU. Планируется усовершенствование модели с целью улучшения точности и адаптации для работы с русскоязычными данными.

Ключевые слова

Машинное обучение, нейронные сети, обработка естественного языка, классификация текстов, предсказание авторства.

Введение

Актуальность и значимость проекта

Современные достижения в области искусственного интеллекта (AI) открывают новые возможности в разных сферах, в том числе и в генерации текстов. Технологии, такие как GPT-3 и другие модели, способны создавать тексты, которые по качеству могут соперничать с человеческими. Это породило важную проблему: как отличить AI-сгенерированный текст от написанного человеком? Особенно это актуально в контекстах, где точность авторства имеет большое значение, таких как академические исследования, журналистика и борьба с дезинформацией.

Цель и задачи проекта

Цель данного проекта заключается в разработке нейросетевой модели, способной предсказывать вероятность того, что текст был сгенерирован искусственным интеллектом. Будет создана модель на основе существующих методов машинного обучения, таких как TF-IDF для векторизации текста и нейросетевых архитектур для классификации. Модель будет обучаться на выборках текстов, как созданных людьми, так и сгенерированных ИИ, с целью достижения высокой точности распознавания.

Описание предметной области

Предметная область проекта охватывает использование машинного обучения в задачах текстовой аналитики, где классификация текстов играет ключевую роль. Использование методов машинного обучения для анализа текстовых данных позволяет выявлять скрытые закономерности и автоматизировать задачи, которые ранее требовали участия человека. В данной работе основное внимание уделяется задаче определения происхождения текста: был ли он создан ИИ или человеком.

Основные результаты работы и их новизна

К текущему моменту времени достигнуты следующие результаты: нейросетевая модель классификации текстов имеет точность 82% на соревновании Kaggle. Также был проведен анализ текущих решений на рынке, и выявлено, что в свободном доступе отсутствуют русскоязычные аналоги для задачи классификации текстов, сгенерированных искусствен-

ным интеллектом. В рамках проекта также будет разработан Telegram-бот, который позволит пользователю в реальном времени отправлять тексты для анализа и получать прогноз вероятности их искусственного происхождения.

Структура работы

Структура работы включает в себя описание проблемы, исследование существующих решений, разработку и обучение модели, а также интеграцию с пользовательским интерфейсом. На следующем этапе будет проведено усовершенствование модели и обучение на русскоязычных данных для повышения точности и расширения области применения.

1 Обзор литературы

1.1 Методы векторизации текста

Один из наиболее распространенных методов представления текстов в числовом формате — TF-IDF (Term Frequency-Inverse Document Frequency), предложенный Salton et al. (1975). Этот метод активно используется для выделения ключевых признаков в текстах, позволяя оценить важность каждого слова относительно всего корпуса документов. Он эффективно применяется в задачах классификации и информационного поиска.

С развитием нейронных сетей появились более продвинутые методы, такие как word2vec (Mikolov et al., 2013) и GloVe (Pennington et al., 2014). Эти подходы формируют плотные векторные представления слов, учитывая их контекст в предложении или документе. Однако для задач классификации текстов, где важно улавливать сложные зависимости и стилевые особенности, простых векторных представлений недостаточно.

1.2 Нейросетевые модели для классификации текстов

Современные исследования показывают, что глубокие нейронные сети, такие как CNN и RNN, обеспечивают более высокую точность классификации текстов по сравнению с традиционными методами. Например, в работе Conneau et al. (2017) использовались двуслойные рекуррентные нейросети, что позволило достичь лучшего качества классификации за счет учета длинных зависимостей в тексте.

В контексте детекции AI-генерированных текстов значительный вклад внесли исследования Zellers et al. (2019), где использовались глубокие нейросетевые модели, обученные

на специально подготовленных датасетах, содержащих как тексты, созданные человеком, так и сгенерированные ИИ. Данное исследование подтвердило возможность высокоточной классификации AI-текстов при наличии достаточного объема обучающих данных.

Кроме того, современные модели трансформеров, такие как BERT (Devlin et al., 2019), показывают выдающиеся результаты в задачах обработки естественного языка. BERT позволяет учитывать контекстное значение каждого слова, что особенно полезно при анализе синтаксической структуры текста и выявлении стилистических особенностей, характерных для AI-генерированных текстов.

1.3 Адаптация методов классификации для русского языка

Большинство существующих исследований сосредоточено на английском языке, и модели, такие как BERT, в первую очередь адаптированы под англоязычные тексты. Для работы с русскоязычными данными требуется дополнительная адаптация, а также использование специализированных русскоязычных моделей, например, RuBERT. Интеграция подобных решений может значительно повысить точность классификации AI-генерированных текстов на русском языке.

Таким образом, использование методов машинного обучения, включая TF-IDF, глубокие нейросети и трансформеры, в сочетании с адаптацией для русскоязычных данных, позволяет эффективно решать задачу детекции AI-генерированных текстов.

2 Описание системы

Система представляет собой нейросетевую модель, предназначенную для предсказания вероятности того, что текст был сгенерирован искусственным интеллектом. Архитектура системы включает несколько ключевых компонентов:

2.1 Модуль загрузки и предобработки данных (Совкова София)

- **Очистка данных:** удаление шума, таких как лишние символы, и приведение текста к стандартному виду.
- **Векторизация текста:** использование метода TF-IDF для преобразования текстов в числовые векторы, подходящие для обработки нейросетью.
- **Разделение на обучающую и тестовую выборки:** данные делятся на две части, одна из которых используется для обучения модели, а другая — для тестирования её качества.

2.2 Модуль нейросетевой классификации (Синицина Софья)

- **Полносвязная нейросеть с тремя слоями:** модель построена на основе трёх полносвязных слоев с активацией ReLU и выходным слоем с функцией активации Sigmoid для предсказания вероятности. (возможны изменения)
- **Обучение модели:** процесс обучения с использованием оптимизатора Adam и функции потерь binary-crossentropy.
- **Оценка точности:** вычисление точности модели на тестовых данных для оценки её эффективности.

2.3 Модуль сохранения и использования модели (Синицина Софья)

- **Сохранение обученной модели:** модель сохраняется в формате .h5 для использования в дальнейшем.
- **Загрузка модели для предсказаний:** возможность загрузки сохраненной модели для выполнения предсказаний на новых текстах.

2.4 Telegram-бот (Совкова София)

- **Взаимодействие с пользователем:** бот получает текстовые сообщения от пользователей через Telegram.
- **Передача текста на анализ:** текст отправляется на сервер, где производится анализ с использованием обученной модели.
- **Вывод результатов:** пользователю предоставляется предсказание с вероятностью того, что текст был сгенерирован ИИ.

3 План дальнейшей работы

3.1 Оптимизация архитектуры модели (Синицина Софья)

В дальнейшем будет проведена оптимизация нейросетевой архитектуры с целью повышения точности модели. Это может включать в себя подбор более сложных слоев, увеличение количества нейронов или использование других архитектур, таких как трансформеры.

3.2 Обучение на русскоязычных данных (Совкова София)

Будет продолжено добавление новых, качественно размеченных данных, с акцентом на русскоязычные тексты. Это позволит улучшить модель в контексте различных языковых особенностей и повысить её общую производительность.

3.3 Дополнение Web-интерфейса (Совкова София)

Для удобства пользователей будет разработан Web-интерфейс (Телеграмм бот), позволяющий взаимодействовать с моделью через браузер или сторонние приложения. Это откроет возможность более широкого использования системы.

3.4 Тестирование системы на больших текстах (Синицина Софья)

Для оценки производительности и стабильности работы модели будут проведены тесты на больших объемах текстовых данных, что позволит выявить потенциальные проблемы в производительности и улучшить модель.

Список литературы

[1] hypertarget1 Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, Lidia S. Chao. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions.

[2] hypertarget2 Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, Cho-Jui Hsieh. Red Teaming Language Model Detectors with Language Models.

[3] hypertarget3 Junchao Wu. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions.

[4] hypertarget4 Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, Yue Zhang. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature.

[5] hypertarget5 Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, Fei Wu. Training-free LLM-generated Text Detection by Mining Token Probability Sequences.

[6] hypertarget6 Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, Yarin Gal. AI models collapse when trained on recursively generated data.