Санкт-Петербургский государственный университет

Программная инженерия

Группа 24.М71-мм

Применение глубокого обучения для обнаружения естественно сгенерированных текстов в русскоязычном контексте

Ван Цзыхань

Отчёт по учебной практике
в форме «Решение»

Научный руководитель:
ст. преподаватель кафедры ИАС, к.ф.-м.н. Азимов Р. Ш.

Санкт-Петербург
2025

Оглавление

Введение

После значительного прорыва в области искусственного интеллекта, крупномасштабные модели, предварительно обученные на больших данных, быстро стали широко распространенными. 30 ноября 2022 года компания OpenAI выпустила чат-бота ChatGPT, основанного на мощной модели GPT 3.5 (Generative Pre-trained Transformer). В отличие от традиционных чат-ботов, ChatGPT завоевал мировое признание благодаря своей способности эффективно понимать контекст, генерировать текст и обладать обширными знаниями. Причина того, что ответы ChatGPT звучат так близко к человеческим, заключается в использовании метода обучения с подкреплением с обратной связью от человека (Reinforcement Learning with Human Feedback, RLHF) [1]. Это как основное преимущество GPT и подобных моделей, так и одна из самых спорных их характеристик.

Несмотря на то, что эти модели продемонстрировали свои мощные возможности в генерации текста, их широкое применение также вызвало множество обсуждений. Билевский Павел Геннадиевич [2] считает, что история искусственного интеллекта полна случаев, когда его потенциал чрезмерно преувеличивался, и такие завышенные оценки обычно использовались для привлечения инвестиций через агрессивный маркетинг. Билевский Павел Геннадиевич выделил несколько потенциальных угроз, связанных с GPT:1.Риски распространения ложной информации и «фейковых новостей»: он утверждает, что работа ChatGPT основана на заранее подготовленных текстовых данных, и ответы генерируются с помощью алгоритмов, при этом эти сгенерированные материалы не имеют указания на источники и не могут быть проверены на «фактическую достоверность».2.Проблемы с авторским правом и авторством: контент, сгенерированный ChatGPT, не имеет четкого автора, что может вызвать вопросы о праве собственности и моральной ответственности. Для ученых это также связано с возможностью академического мошенничества.3.Чрезмерная зависимость от технологий: излишняя зависимость от искусственного интеллекта и уменьшение роли человека могут при-

вести к тому, что люди не смогут сохранять ведущую роль в процессе технологического развития.

В отличие от предыдущих утверждений, Людмила Анатольевна Иванова [3] занимает более сбалансированную позицию. Она отмечает, что технологии, такие как GPT, обладают существенными преимуществами, в частности, в повышении эффективности написания текстов. Однако, по её мнению, использование искусственного интеллекта должно быть сопряжено с соблюдением строгих ограничений и стандартов, что позволит минимизировать возможные риски и обеспечить этическую корректность применяемых решений.

В настоящее время уже существует множество соответствующих технологий, таких как GPTZero [4], Originality.AI [5], DetectGPT [6], которые используют методы глубокого обучения для обнаружения искусственно сгенерированных текстов. В данной работе будет проведен анализ сильных и слабых сторон основных нейронных сетевых моделей, а также проведены улучшения для модели, ориентированной на русский язык, в области обнаружения, и в конечном итоге модель будет упакована в исполнимую программу.

1. Постановка задачи

Целью данной работы является оптимизация существующей модели и реализация генерации естественного языка на основе русского языка. Для выполнения этой цели были поставлены следующие задачи, которые будут решаться с использованием языка программирования Python:

1. Собрать датасет и классифицировать его на тексты, написанные человеком, машинные тексты и смешанные тексты.

2. Исследовать модели DetectGPT и GPTzero, проанализировать их стратегии и методы, применяемые для обнаружения машинно-сгенерированных текстов.

3. Провести тонкую настройку модели RuBERT, определив стратегии подбора характеристик для улучшения точности и производительности.

4. На основе результатов предыдущих экспериментов разработать эффективную гибридную систему обнаружения, оптимизируя архитектуру модели для повышения производительности и уменьшения её сложности.

5. Упаковать оптимизированную модель в исполнимое приложение и разработать простой пользовательский интерфейс.

2. Обзор

Существующие методы обнаружения можно условно разделить на два типа:(1) Методы обнаружения на основе признаков(2) Методы обнаружения на основе глубокого обучения[7].Основное различие между ними заключается в способах извлечения признаков и возможностях модели.

2.1. Методы обнаружения на основе признаков

Первая группа методов обнаружения основана на извлечении и анализе различных статистических характеристик машинно-сгенерированных текстов для определения их источника (то есть, человек или машина). Эти характеристики обычно связаны с языковыми паттернами текста, синтаксической структурой, использованием лексики, генерационными закономерностями и т. д. Одним из существующих моделей является GLTR (Giant Language model Test Room) [8], который использует ряд статистических методов на основе признаков для обнаружения машинно-сгенерированных текстов. Конкретно, его ключевая технология включает три теста: вероятность генерации слов, их ранжирование и энтропия контекста. Эти характеристики выявляют особенности машинно-сгенерированных текстов — они обычно сосредоточены на словах с высокой вероятностью и слишком уверены в генерации слов в условиях низкой энтропии. Эксперименты показали, что GLTR значительно повышает точность распознавания фальшивых текстов пользователями, с 54% без использования инструмента до более 72% при использовании. Однако у GLTR есть свои ограничения, особенно при работе с враждебно сгенерированными текстами и новыми генеративными моделями, его эффективность может снизиться. Kristina Schaaff и другие [9] обобщили часто используемые текстовые признаки, которые включают восемь основных категорий, таких как сложность (Perplexity), семантика (Semantic), читаемость (Readability), текстовый вектор (Text Vector) и т. д. Заключения экспериментов показали, что текстовые векторы, извлечённые с использованием предобученной модели BERT, и

структурные характеристики документа играют решающую роль в обучении модели. Методы обнаружения на основе статистики текстовых признаков сильно зависят от выбора признаков, что требует высокого уровня лингвистических знаний от исследователей, что затрудняет их распространение. Это привело к развитию методов обнаружения на основе глубокого обучения.

2.2. Методы обнаружения на основе нейронных сетей

Методы обнаружения на основе глубокого обучения сначала требуют предварительной обработки текста, затем преобразуют текст в вектор слов, а затем с помощью глубоких нейронных сетей извлекают характеристики текста, а не вручную отбирают их, что значительно увеличивает допустимость ошибок. ZhiWu Fan и другие [10] исследовали традиционные методы, такие как TextCNN, TextRNN, а также активные в области обработки естественного языка модели, такие как Transformer и DPCNN, и обнаружили, что использование глубокой пирамидальной свертки (DPCNN) обеспечивает высокую точность на уровне 96,85%. Кроме того, Mitchell и другие [11] предложили модель для обнаружения без обучающих примеров — DetectGPT. DetectGPT — это метод обнаружения без обучающих примеров, основанный на кривизне вероятности, предназначенный для выявления текста, сгенерированного большими языковыми моделями (LLM). Он не требует дополнительных обучающих данных, отдельного классификатора или технологии водяных знаков текста, а использует анализ вероятностной структуры LLM и возмущений текста для выполнения обнаружения. Исследования показали, что текст, сгенерированный машиной, и текст, написанный человеком, имеют различия в кривизне распределения вероятности, особенно в области отрицательной кривизны. Тексты, написанные человеком, обычно не подвергаются значительным изменениям после добавления небольших возмущений, в то время как машинное обучение оказывает значительное влияние на изменения после возмущений. Результаты показали, что

DetectGPT всегда обеспечивал наивысшую точность AUROC в задаче обнаружения на нескольких наборах данных (таких как XSum, SQuAD и WritingPrompts) и на нескольких моделях (например, GPT-2, Neo-2.7, NeoX и др.), что еще раз подчеркивает преимущества глубокого обучения по сравнению с традиционными методами машинного обучения для поиска признаков.

Однако текущие модели недостаточно адаптированы для русского языка. Например, для моделей DetectGPT и GPTzero ниже мы покажем три изображения, иллюстрирующих реальные примеры.

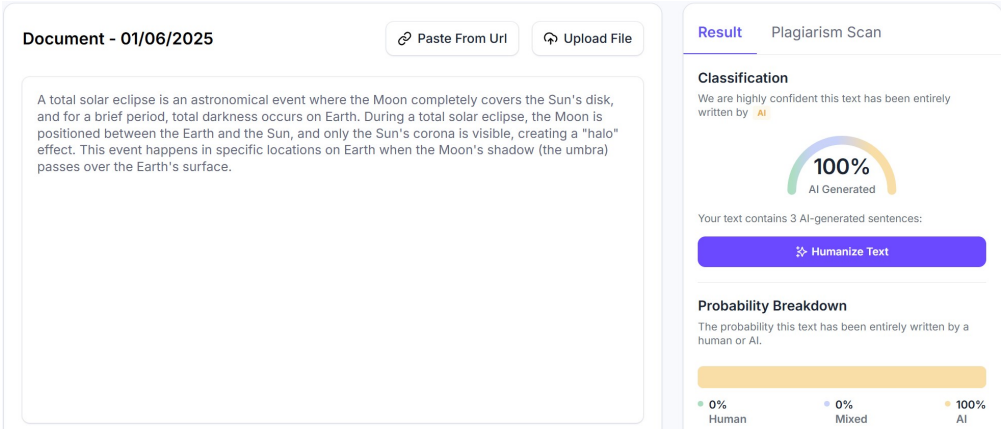

Рис. 1: Эффективность модели DetectGPT в контексте английского языка

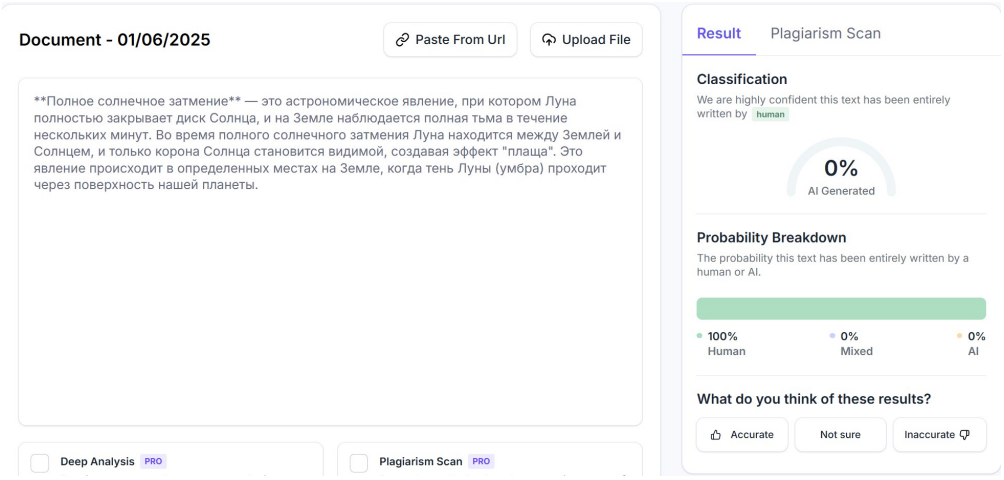

Рис. 2: Эффективность модели DetectGPT в контексте русского языка

На первом изображении модель успешно идентифицировала, что английский текст, сгенерированный GPT, был произведен искусственным интеллектом. Однако на втором изображении модель ошибочно

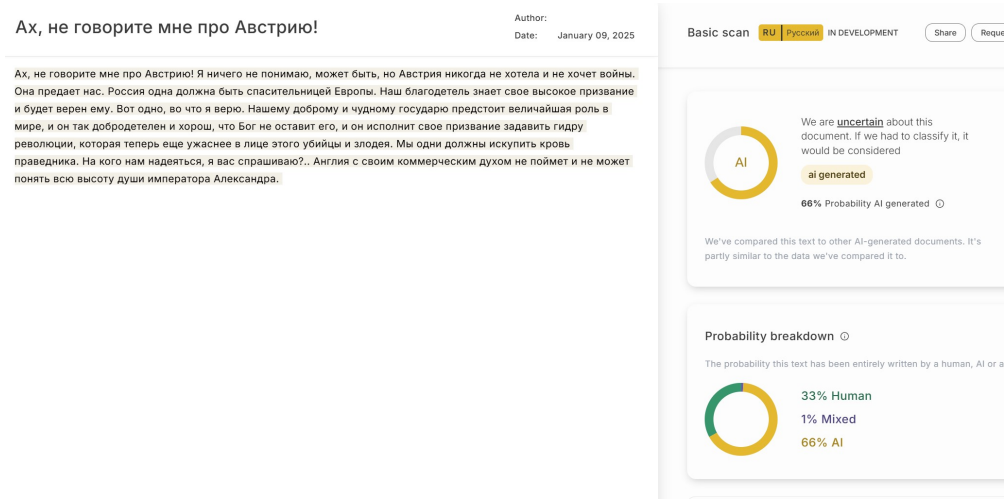

Рис. 3: Эффективность модели GPTzero в контексте русского языка

оценила вероятность того, что русский текст был сгенерирован ИИ, как 0%, что является недостоверным выводом. На третьем изображении, при использовании модели GPTzero для анализа фрагмента из ”Войны и мира”, система оценила вероятность того, что этот текст был сгенерирован ИИ, на уровне 66%, что также представляет собой ошибочную интерпретацию.

3. Описание решения

3.1. Сбор данных

Сбор данных является ключевым элементом исследовательского процесса, качество которого напрямую определяет надежность и эффективность результатов эксперимента. Для обеспечения высокого качества набора данных планируется использовать комбинированный подход, включающий различные методы.

Для анализа текстов будет осуществляться извлечение данных из различных источников. На первом этапе планируется использовать существующие наборы вопросов и ответов, такие как диалоговые материалы, представленные в мессенджерах. Далее источниками станут авторитетные новостные издания, включая телеканалы и их цифровые платформы. Учитывая динамику изменений языковых норм и практик под воздействием времени, также будут привлечены современные художественные произведения.

Основным методом получения машинного текста является разработка подсказок для моделей и массовая генерация текстовых данных с использованием API GPT. Для повышения способности модели к обобщению часть текстов, созданных человеком, подвергается повторной генерации с использованием крупной языковой модели, что приводит к формированию смешанных текстовых данных.

3.2. Исследование технологий DetectGPT и GPTzero

Основная гипотеза DetectGPT заключается в том, что машинно сгенерированный текст на логарифмическом уровне вероятности обычно находится ближе к максимальному значению логарифмической вероятности, чем текст, написанный человеком, и он более чувствителен к возмущениям текста. Различие в источнике текста можно определить, вычислив разницу логарифмических вероятностей между исходным

текстом и текстом с добавленными возмущениями.Формула следующая:

$$d(x, p_\theta, q) = \log p_\theta(x) - \mathbb{E}_{\tilde{x} \sim q(\cdot|x)} \log p_\theta(\tilde{x})$$

где:

- $\log p_\theta(x)$: представляет прогнозируемую вероятность выбора текста x согласно генеративной модели.
- $\mathbb{E}_{\tilde{x} \sim q(\cdot|x)} \log p_\theta(\tilde{x})$: представляет среднюю прогнозируемую вероятность для всех возмущённых текстов $\tilde{x}$ согласно генеративной модели.
- $d(x, p_\theta, q)$: выражает разницу между прогнозируемыми вероятностями исходного текста и возмущённого текста, используется для оценки того, находится ли текст в области с отрицательной кривизной.

Если $d(x, p_\theta, q)$ велико, это означает, что текст, скорее всего, сгенерирован моделью p_θ; если мало — текст, вероятно, создан человеком.

Ключевая гипотеза GPTZero заключается в том, что машинно сгенерированный текст демонстрирует специфические статистические свойства, такие как низкая сложность (перплексия) и высокая последовательность вероятности токенов. GPTZero использует два основных параметра для оценки текста: перплексию (Perplexity) и вероятность всплесков (Burstiness). Перплексия (Perplexity):

$$PPL(x) = 2^{-\frac{1}{N} \sum_{i=1}^{N} \log_2 p(x_i)}$$

где:

- N — общее количество токенов в тексте.
- $p(x_i)$ — вероятность i-го токена в тексте x.

Рывки вероятности (Burstiness):

$$B(x) = \frac{1}{N-1} \sum_{i=2}^{N} |p(x_i) - p(x_{i-1})|$$

где:

- $p(x_i)$ и $p(x_{i-1})$ — вероятности текущего и предыдущего токена.

Если $PPL(x)$ низко и $B(x)$ также низко, текст, скорее всего, является машинно сгенерированным. Если же $B(x)$ велико, текст, вероятно, создан человеком. Эти два показателя используются совместно для увеличения точности определения источника текста.

3.3. Доработка RuBERT

BERT внес в архитектуру Transformer важное изменение, внедрив двусторонний кодировщик, который с помощью двух задач: Masked Language Model (MLM) и Next Sentence Prediction (NSP), позволяет модели извлекать глубокие семантические связи как из предыдущего, так и из следующего контекста. RuBERT основан на архитектуре BERT и представляет собой предобученную языковую модель, оптимизированную специально для русского языка.

В задаче текстовой классификации можно улучшить производительность модели, сочетая статистические признаки, такие как различия в возмущении, перплексия и всплесковость, с семантическими признаками двумя способами. Первый способ заключается в использовании статистических признаков, таких как $d(x, p_\theta, q)$, перплексия и всплесковость, для управления обучением модели. Второй способ — это прямое сочетание статистических признаков и семантических вложений. Извлекаются статистические признаки, такие как $d(x, p_\theta, q)$, перплексия и всплесковость, и комбинируются с вложениями предложений RuBERT (вектор $[CLS]$), чтобы сформировать полный вектор признаков, который затем вводится в нейронную сеть классификатора для классификации. Конкретные стратегии требуют дальнейшей экспериментальной проверки.

Заключение

Сделано: Проект уже получил одобрение научного руководителя — Азимова Рустама Шухратулловича. Проведён анализ научных статей, что позволило определить основные этапы реализации проекта.

Планы: В весенний семестр завершить процесс сбора данных, необходимых для работы. Провести воспроизведение фреймворков DetectGPT и GPTZero, которые являются основой исследования.

Список литературы

[1] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback [EB]. arXiv:2203.02155.

[2] Былевский, П. Г. (2023). Культурологическая деконструкция социально-культурных угроз ChatGPT информационной безопасности российских граждан. Философия и культура, 8.

[3] Иванова, Л. А. (доктор педагогических наук, доцент). (2024). Московский государственный технический университет гражданской авиации (Иркутский филиал). Коммунаров, 3, Иркутск, 664047, Россия.

[4] "GPTZero," [Online]. Available: https://en.wikipedia.org/wiki/GPTZero

[5] Originality.AI, [Online]. Available: https://originality.ai/

[6] "DetectGPT," [Online]. Available: https://detectgpt.com/

[7] Crothers, E. N., Japkowicz, N., & Viktor, H. L. (2023). Machine-generated text: A comprehensive survey of threat models and detection methods. IEEE Access, 11, 70977-71002.

[8] Gehrmann, S., Strobelt, H., & Rush, A. M., "GLTR: Statistical detection and visualization of generated text," arXiv preprint arXiv:1906.04043, 2019. [Online].

[9] K. Schaaff, T. Schlippe, and L. Mindner, "Classification of human and AI-generated texts for English, French, German, and Spanish," Proc. Int. Conf. on Natural Language and Speech Processing, 2023, pp. 1-10.

[10] Zhiwu, F., & Jinliang, Y. (2024). "Text detection method for ChatGPT-generated text based on deep pyramid convolutional neural network," Data Analysis and Knowledge Discovery, 8(7),

[11] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature," arXiv preprint, arXiv:2301.11305, 2023.