Детектирование искусственно сгенерированного научного текста

Стельмах Т.Д., СПбГУ, Санкт-Петербург tanya252002@gmail.com

Аннотация

В рамках данной работы рассматривается проблема детектирования искусственно сгенерированных научных текстов, что важно для поддержания достоверности и качества научных публикаций, а также защиты академической честности. Существующие методы детектирования имеют ограничения, особенно в отношении устойчивости к перефразированию. В данной работе предлагается использование методов топологического анализа данных (TDA) для детектирования таких текстов, которые показали высокую точность и устойчивость. Был создан датасет и проведено тестирование методов на реальных и сгенерированных текстах, что позволило выявить преимущества предложенных подходов.

Введение

В последние годы технологии искусственного интеллекта, такие как популярные модели ChatGPT (для генерации текста), Sora (для генерации видео) и Suno (для генерации музыки), становятся всё более доступными и широко используемыми. Эти технологии оказывают значительное влияние на различные сферы нашей жизни, от развлечений и медиа до науки и образования.

Подробнее рассмотрим научную область. Языковые модели способны быстро и убедительно генерировать научные тексты, включая ложные факты. Кроме того, такие тексты могут искусственно повышать рейтинг цитируемости отдельных научных изданий и увеличивать количество публикаций у отдельных учёных, что ведёт к искажению научной репутации и обесцениванию научных достижений. Это также может широко использоваться в студенческих и школьных работах, что приводит к снижению качества образования. Например, студенты могут использовать искусственно сгенерированные тексты для написания курсовых и дипломных работ, не приобретая при этом реальных знаний и навыков.

В связи с этим, в рамках текущей работы мы будем рассматривать проблему детектирования искусственно сгенерированных научных текстов, в частности, статей. Это исследование важно для поддержания достоверности и ка-

чества научных публикаций, а также для защиты академической честности и предотвращения распространения дезинформации.

Существующие методы детектирования сгенерированных текстов имеют значительные ограничения, особенно в отношении устойчивости к перефразированию. Например, при задании модели контекста, в котором она должна сгенерировать текст, имитирующий человеческий, многие детекторы оказываются неэффективными. Это особенно критично в условиях, когда генеративные модели становятся всё более продвинутыми и способны создавать тексты, практически неотличимые от написанных человеком.

Для иллюстрации можно рассмотреть детектор DetectGPT. Как видно на рисунке 1, этот детектор плохо справляется с перефразированными текстами, иногда даже давая нулевую вероятность того, что текст был сгенерирован.

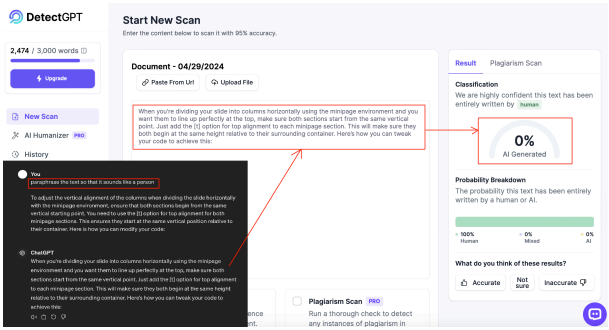

Рис. 1: Пример работы детектора DetectGPT

Таким образом, существует острая необходимость в разработке детектора, который был бы устойчив к перефразированию и изменению запросов (prompt) модели.

Обзор методов

Существует множество подходов к детектированию искусственно сгенерированных текстов. Рассмотрим основные из них:

1. **Статистические методы**: Статистические методы анализируют различные статистические особенности текста, такие как частоты слов (частотный анализ n-грамм), длины предложений и словарный запас [1]. Например, модели могут выявлять аномалии в распределении этих признаков, которые указывают на искусственное происхождение текста. Такие методы могут быть достаточно эффективными, но их точ-

ность может снижаться при работе с текстами, которые были специально перефразированы для обхода детекторов.

2. **Методы глубокого обучения**: Современные методы глубокого обучения, такие как BERT или GPT, могут быть дообучены для распознавания признаков искусственного текста [5]. Эти модели анализируют последовательности слов и контекстуальные связи, что позволяет им выявлять тонкие отличия между реальными и сгенерированными текстами. Методы глубокого обучения показывают высокую точность, но требуют значительных вычислительных ресурсов для обучения и работы.

3. **Измерение размерности подпространства представления данных**: Этот метод основывается на идее, что данные часто образуют многообразие меньшей размерности, чем всё пространство, в котором они представлены [6]. Мы стараемся измерить эту размерность, используя модели типа RoBERTa. Для этого необходимо взять токенизатор, пропустить через него данные, а затем подать их на вход модели для получения эмбеддингов. Далее анализируется вектор первого токена, который агрегирует информацию всего текста. Этот метод позволяет выявлять структурные особенности текста, указывающие на его искусственное происхождение.

Метод

В ходе изучения научных статей и материалов, а также проведённых самостоятельных экспериментов, было принято решение использовать методы топологического анализа данных (TDA) для детектирования искусственно сгенерированных текстов. Методы TDA показали высокую точность и устойчивость по сравнению с традиционными статистическими методами [6].

Основные преимущества методов PHD и MLE:

- **Универсальность**: Способны детектировать тексты, сгенерированные различными моделями, включая новые модели.
- **Робастность**: Более устойчивы к шуму по сравнению с другими методами, связанными с подсчётом топологических размерностей.
- **Устойчивость к перефразированию**: Метод PHD особенно хорошо справляется с перефразированными текстами.

- **Стабильность при смене темы текста**: Методы остаются эффективными независимо от тематики текста.

Остальные методы не попали в наше сравнение, поскольку либо они слишком ресурсозатратны, либо их метрики значительно хуже.

Создание датасета

Для детектирования искусственно сгенерированных научных статей потребовалось создать собственный датасет, так как подходящего готового датасета с научными статьями не нашлось.

Наиболее доступными и эффективными англоязычными языковыми моделями, согласно leaderboard'ам, оказались ChatGPT 3.5 (относительно дешевое API) и модели семейства LLaMA (бесплатные). Эти модели использовались для генерации текста, подавая на вход предложение, написанное человеком, и продолжая его. Тексты в среднем не превышали двух-трёх абзацев.

В качестве частей для детектирования были выбраны Abstract и Introduction. Abstract является кратким изложением содержания статьи, а Introduction — введением, которое часто не несет значительной смысловой нагрузки. Это делает эти части особенно подверженными генерации текста.

Темы статей были связаны с программированием, анализом данных и машинным обучением, так как специалисты в этих областях обычно лучше осведомлены о новинках ИИ. Важно было выбирать статьи, написанные до расцвета LLM (до 2018 года). Эти темы и ключевые слова для поиска статей были сгенерированы языковой моделью, и по каждой теме было найдено не менее 20 ключевых слов для поиска на Google Scholar.

Для парсинга статей был использован старый парсер 2015 года [7], который обновили и модифицировали для извлечения полного текста статьи. Поскольку многие PDF-ридеры некорректно восстанавливают текст, было принято решение использовать языковую модель для обработки полученного текста.

Результаты

Результаты измерения внутренней размерности показали, что научные тексты имеют большую внутреннюю размерность по сравнению с текстами из Википедии [6]. Это связано с более специализированной лексикой и сложными структурами предложений в научных статьях. Также диапазон внутренней

размерности для настоящих научных текстов оказался шире, что объясняется их сложностью и разнообразием тем и стилей написания.

Также было установлено, что новые модели, такие как LLaMA 3 70B, генерируют тексты, более похожие на человеческие. Это связано с тем, что они лучше улавливают нюансы научного языка, что приводит к меньшим различиям в оцененных размерах между сгенерированными и реальными текстами. Методы MLE и PHD показали различную чувствительность к сложности текстов. PHD продемонстрировал большую чувствительность к сложным структурам научных статей, что способствовало более явному разделению между сгенерированными и реальными текстами.

Metric	**PHdim SVM**	**PHdim GB**	**MLE SVM**	**MLE GB**
Accuracy	0.59	0.59	0.60	0.59
Precision (Gen)	0.82	0.87	0.74	0.64
Precision (Human)	0.55	0.55	0.56	0.58
Recall (Gen)	0.23	0.21	0.30	0.46
Recall (Human)	0.95	0.97	0.89	0.74
F1-score (Gen)	0.36	0.34	0.43	0.54
F1-score (Human)	0.70	0.70	0.69	0.65
Validation Accuracy	0.60	0.59	0.57	0.57
ROC-AUC	0.59	0.58	0.55	0.55

Таблица 1: Результаты

В ходе сравнения метрик было замечено улучшение практически по всем важным нам показателям по сравнению с результатами, полученными с помощью датасета со статьями из Википедии [6]. Особое внимание уделялось показателю Precision (Gen), который также улучшился. Метод PHD в этом тестировании показал значительно лучшие результаты и высокую точность.

Список литературы

[1] Gehrmann, S., Strobelt, H., and Rush, A.M., 2019. *GLTR: Statistical Detection and Visualization of Generated Text*. arXiv preprint arXiv:1906.04043.

[2] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y., 2019. *Defending against Neural Fake News*. Advances in Neural

Information Processing Systems, 32. Available at: https://arxiv.org/abs/1905.12616.

[3] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T., 2023. *A Watermark for Large Language Models*. arXiv preprint arXiv:2301.10226. Available at: https://arxiv.org/abs/2301.10226.

[4] Krishna, K., Song, Y., Karpinska, M., Wieting, J., and Iyyer, M., 2023. *Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval is an Effective Defense*. arXiv preprint arXiv:2303.13408. Available at: https://arxiv.org/abs/2303.13408.

[5] Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., and Wu, Y., 2023. *How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection*. arXiv preprint arXiv:2301.07597. Available at: https://arxiv.org/abs/2301.07597.

[6] Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Piontkovskaya, I., Nikolenko, S., and Burnaev, E., 2023. *Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts*. arXiv preprint arXiv:2306.04723. Available at: https://arxiv.org/abs/2306.04723.

[7] fjxmlzn, 2023. *Scholar*. GitHub repository. Available at: https://github.com/fjxmlzn/scholar.