DOI 10.37882/2223-2966.2024.06.40

ИССЛЕДОВАНИЕ МЕТОДОВ И РЕСУРСОВ ОБНАРУЖЕНИЯ ТЕКСТОВЫХ БЛОКОВ ИНФОРМАЦИИ, СГЕНЕРИРОВАННЫХ ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ: НОВЫЕ ПОДХОДЫ И ПЕРСПЕКТИВЫ

RESEARCH OF METHODS AND RESOURCES FOR DETECTING TEXT BLOCKS OF INFORMATION GENERATED BY ARTIFICIAL INTELLIGENCE: NEW APPROACHES AND PERSPECTIVES

Ya. Ukuahamba

Summary. In the modern information society, where the data flow is becoming more extensive and complex, the issue of detecting information recreated by generative systems is becoming relevant and significant. ChatGPT is based on machine learning, which is currently the most popular method in artificial intelligence (AI) technology. One of the key aspects is human interaction with generative systems that create, modify or adapt information. These systems include machine learning and an inexhaustible amount of knowledge and information, on the basis of which the user of the system is given the opportunity to create and modify information with a high degree of realism. The average person, for example, is increasingly becoming a victim of fake information generated by neural network models. The article discusses the resources for detecting text blocks of information created by humans during interaction and correction by a generative system.

Keywords: artificial intelligence, ChatGPT, OpenAI, information, language model, plagiarism.

Укуахамба Ядмилде Авелину
аспирант, Институт инженерных и цифровых технологий, г. Белгород
1027419@bsu.edu.ru

Аннотация. В современном информационном обществе, где поток данных становится все более обширным и сложным, вопрос об обнаружении информации воссозданной генеративными системами становится актуальным и значимым. ChatGPT основан на машинном обучении, которое в настоящее время является самым популярным методом в технологии искусственного интеллекта (ИИ). Одним из ключевых аспектов является взаимодействие человека с генеративными системами, которые создают, изменяют или адаптируют информацию. Данные системы включают в себя машинное обучение и неисчерпаемый объём знаний и информации, на основе которого пользователю системы предоставляется возможность создания и модификации сведений с высокой степенью реализма. Среднестатистический человек, к примеру, всё чаще становится жертвой поддельной информации, сгенерированный моделями нейронных сетей. В статье рассмотрены ресурсы обнаружения текстовых блоков информации, созданных человеком при взаимодействии и коррекции генеративной системой.

Ключевые слова: искусственный интеллект, ChatGPT, OpenAI, информация, языковая модель, плагиат.

Развитие технологий искусственного интеллекта и его влияние на многие сферы жизни человека в последние годы стало темой, вызывающей растущую озабоченность. Системы искусственного интеллекта нового поколения, такие как чат-боты, стали более доступными в Интернете и более мощными с точки зрения возможностей.

В контексте информационных технологий, передовые модели искусственного интеллекта, такие как ChatGPT, GPT-3.5, BERT, RoBERTa и XLNet, являются значительным прорывом. Эти модели, разработанные гигантами технологической индустрии, такими как OpenAI, Google и Microsoft, предоставляют целый ряд преимуществ, включая улучшенное пользовательское взаимодействие, сотрудничество и доступность.

Модель ChatGPT и ее представители 3-го поколения предобученных генеративных моделей, выпущенных компанией OpenAI, являются одной из наиболее сильных генеративных моделей, существующих в настоящее время. Большие языковые модели (LLM) — это передовые системы искусственного интеллекта (ИИ), предназначенные для обработки, понимания и создания текста, подобного человеческому. Примерами таких языковых моделей являются BERT, RoBERTa и другие. Они основаны на методах глубокого обучения и обучены на массивных наборах данных, обычно содержащих миллиарды слов из различных источников, таких как веб-сайты, книги и статьи [Былевский, 2023; Давлетов, 2023; Зонова, 2023; Ивахненко, 2023].

ChatGPT — это тип нейронных языковых моделей, впервые представленных компанией OpenAI, которые обучаются на больших наборах текстовых данных, чтобы генерировать текст, схожий с человеческим. Предобучение относится к начальному процессу обучения

на корпусе, в результате которого модель учится предсказывать следующее слово в тексте и получает основу для успешного выполнения дальнейших задач, не имея больших объёмов данных. GPT являются «трансформерами», которые представляют собой тип нейросетей, использующих механизм самосвязываемости для обработки последовательных данных. Они могут быть дообучены для различных задач обработки естественного языка, таких как генерация текста, машинный перевод и классификация текста [Ладыжец, 2023].

Как отмечает автор и критикует ChatGPT: «он лишен разумного человеколюбия, не мыслит и вообще неразумно (исходя из философского определение интеллекта)» [Hanna, 2023].

Обор современных генеративных моделей

ChatGPT — чат-бот с генеративным искусственным интеллектом, разработанный компанией OpenAI и способный работать в диалоговом режиме, поддерживающий запросы на естественных языках. Система способна отвечать на вопросы, генерировать тексты на разных языках, включая русский, относящиеся к различным предметным областям. Важной особенностью является возможность генерации по запросу программ на различных языках программирования [Лёвин, 2022; Малышев, Смирнов, 2024; Намиот, Зубарева, 2023].

К основным типам генеративных моделей, которые широко используются в современных генеративных алгоритмах, относятся: генеративно-состязательные сети (GAN), вариационные автоэнкодеры (VAE), и модели на основе трансформаторов. Трансформаторы — преобразователи, которые представляют собой усовершенствованный тип архитектуры нейронной сети, широко используемый в исследованиях LLM. Ниже представлена схема основных генеративных моделей (рисунок 1).

Разработка инструментов обнаружения плагиата на основе искусственного интеллекта подтверждается фактическими данными. Действительно, с момента запуска ChatGPT список интернет-ресурсов для инструментов и сервисов обнаружения контента, созданных ИИ, растет буквально еженедельно.

Текст можно быстро проанализировать и сравнить с помощью систем обнаружения плагиата на базе искусственного интеллекта, чтобы обнаружить случаи плагиата. Эти инструменты используют алгоритмы для обнаружения сходства между представленными работами и обширной базой данных источников, помогая более эффективно выявлять потенциальные случаи плагиата.

Исследование оригинальности текстов, сгенерированных ИИ

Существуют некоторые исследования и инструменты, которые занимаются вопросами оригинальности текстов, сгенерированных искусственным интеллектом. К таким можно отнести GPTZero — инструмент, который анализирует текст и определяет, написан ли он человеком или искусственным интеллектом, выделяя конкретные предложения, которые предположительно были созданы искусственным интеллектом [Оравек, 2023: 217]. Raidar — это метод, который позволяет определить, написан текст человеком или создан искусственным интеллектом, не имея доступа к внутреннему устройству модели [Alkaissi, McFarlane, 2023]. Raidar сравнивает оригинальный текст с переписанным и определяет количество изменений. Также сюда можно отнести Copyleaks — универсальный инструмент для проверки подлинности

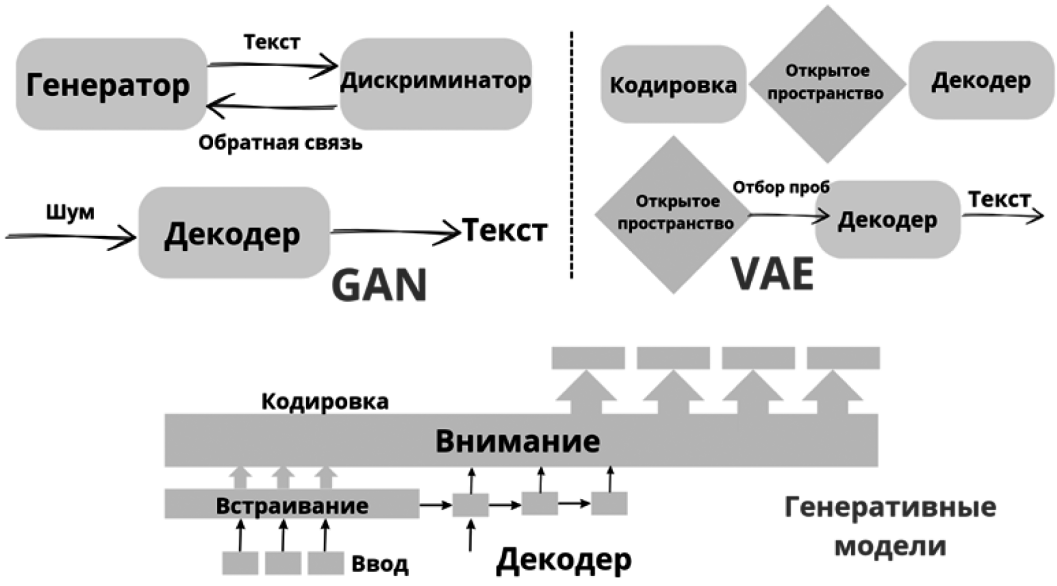

Рис. 1. Схема основных генеративных моделей

текста, сгенерированного искусственным интеллектом, и поддержания целостности контента [Kobis, 2021].

Несмотря на то, что эффективность различных инструментов и методов искусственного интеллекта может отличаться, особое внимание следует уделить тому, как они взаимодействуют с короткими текстами. Большинство исследований и инструментов обычно сосредоточены на обработке более длинных текстов, и их эффективность при работе с короткими текстами может быть не такой высокой.

В эпоху цифровой трансформации и развития искусственного интеллекта возникает вопрос о способности ИИ генерировать оригинальный контент. Особый интерес представляет то, как системы проверки на плагиат реагируют на короткие тексты, сгенерированные ИИ.

Методика обнаружения коротких текстовых блоков информации

Исследовательская гипотеза предполагает, что короткие тексты объемом до 100 слов, сгенерированные ИИ, не будут распознаны системами проверки на плагиат. Это может быть связано с тем, что ИИ способен создавать уникальные комбинации слов и фраз, которые отличаются от уже существующих текстов. Кроме того, короткие тексты могут не содержать достаточного количества информации для сопоставления с большими базами данных, предлагаемыми системами проверки.

Цель исследования заключалась в оценке вероятности верной классификации искусственно порожденного текстового контента в контексте взаимодействия генеративных систем с человеком. Объектом вступают методы оценки систем выявления воссозданного контента. Предметом — текстовый контент.

H_0: текст полностью принадлежит человеку.

H_1: текст содержит сгенерированные блоки.

Матрица ошибок представляет собой эффективный инструмент для анализа эффективности методов обнаружения и классификации блоков текста, сгенерированных искусственно. В частности, «ложные положительные» результаты (ошибки первого рода) в матрице ошибок могут указывать на случаи, когда системы антиплагиата неверно классифицируют сгенерированные блоки текста как созданные человеком. Это подтверждает наблюдение о том, что системы антиплагиата не обнаруживают блоки текста, сгенерированные генеративной системой и состоящие из менее 100 слов.

Подобно тому, «ложные отрицательные» результаты (ошибки второго рода) могут свидетельствовать о ситуациях, когда текст, созданный человеком, неправильно классифицируется как сгенерированный. Это может быть связано с обнаружением стеганографии в текстах, созданных искусственным интеллектом.

Для сравнения прогнозов с реальностью используется матрица ошибок — таблица, содержащая четыре различные комбинации прогнозируемых и фактических значений. Прогнозируемые значения описываются как

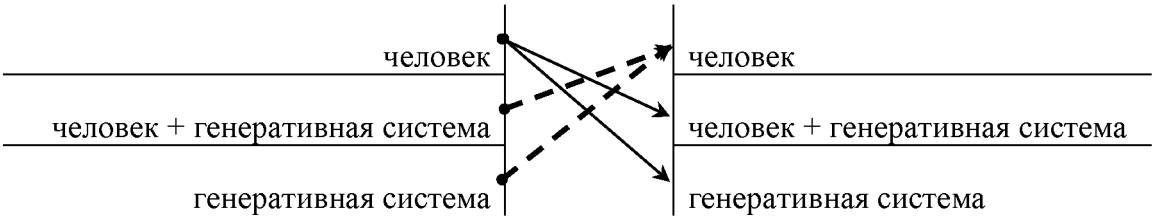

Таблица 1

Матрица ошибок

H_0: текст полностью принадлежит человеку.		Результат анализа	
		Все написанные блоки принадлежат человеку (**гипотеза принимается**)	обнаружены сгенерированные блоки (**гипотеза отвергается**)
Автор	написано человеком (**верна**)	все элементы текста принадлежат человеку	Ошибочно классифицированные блоки ($P_{\varnothing}$, ошибка I-го рода)
	имеются сгенерированные блоки (**не верна**)	текст ошибочно принят за написанный человеком ($P_{\varnothing}$, ошибка II-го рода)	в тексте встречаются сгенерированные блоки

положительные и отрицательные, а фактические значения описываются как истинные и ложные. Обычно матрица ошибок используется для оценки точности моделей в задачах классификации. Однако прогнозирование и распознавание образов можно считать частным случаем этой проблемы, поэтому матрица ошибок также важна для измерения точности прогнозирования.

Современные нейронные сети уже достаточно развиты для создания высококачественных текстов, которые сложно отличить от текстов, созданных человеком. Однако в большинстве своем люди способны понять, что сгенерированные тексты принадлежат нейронным сетям, благодаря примитивности создаваемых объектов, ошибкам и неточностям. Написание текстов попрежнему остается сложной задачей для искусственного интеллекта, поскольку человеческий язык является более сложной областью для моделирования и он имеет более высокий уровень абстракции и контекстуальности.

Заключение

Уникальный алгоритм системы проверки на плагиат, позволяющий автоматически выявлять и отмечать машинно сгенерированные фрагменты текста, является эксклюзивной разработкой компании «Антиплагиат», созданной на основе многолетних исследований в области обработки естественного языка. Система проверки сканирует текст на предмет фрагментов, предположительно созданных с использованием моделей глубокого обучения GPT2, GPT-3, ChatGPT. Детектор машинного текста обучен и протестирован на большом наборе данных, что сводит к минимуму риск ложных срабатываний. Если в проверяемом документе обнаруживаются сгенерированные фрагменты, система антиплагиата выделяет их, а сам документ помечается как «Подозреваемый».

При проведенном исследовании, где анализировался короткий текст (до 100 слов), система проверки на плагиат не обнаружила текст, сгенерированный нейросетью. И как указывает разработчик, детектор текстов обучен и протестирован на большом наборе данных, но результаты исследования показали, что существует непосредственная связь между количеством слов в проверяемом тексте. В дальнейшем, чтобы повысить надежность результатов исследования, необходимо принимать более крупные выборки.

Рис. 2. Результаты проверки

ЛИТЕРАТУРА

1. Былевский П.Г. Культурологическая деконструкция социальнокультурных угроз ChatGPT информационной безопасности российских граждан // Философия и культура. 2023. №8. URL: https://cyberleninka.ru/article/n/kulturologicheskaya-dekonstruktsiya-sotsialnokulturnyh-ugroz-chatgpt-informatsionnoy-bezopasnosti-rossiyskih-grazhdan
2. Давлетов А.Р. Главные трудности при интеграции машинного обучения в коммерческую эксплуатацию // Инновации и инвестиции. 2023. №10. URL: https://cyberleninka.ru/article/n/glavnye-trudnosti-pri-integratsii-mashinnogo-obucheniya-v-kommercheskuyu-ekspluatatsiyu
3. Зашихина И.М. Подготовка научной статьи: справится ли ChatGPT? // Высшее образование в России. 2023. №8-9. URL: https://cyberleninka.ru/article/n/podgotovka-nauchnoy-stati-spravitsya-li-chatgpt
4. Зонова Д.Ю. ПРИНЦИП РАБОТЫ И ПРОБЛЕМЫ «GENERATIVE PRE-TRAINED TRANSFORMER ARTIFICIAL INTELLIGENCE» // Вестник науки и образования. 2023. №8 (139). URL: https://cyberleninka.ru/article/n/printsip-raboty-i-problemy-generative-pre-trained-transformer-artificial-intelligence
5. Ивахненко Е.Н., Никольский В.С. ChatGPT в высшем образовании и науке: угрозы или ценный ресурс? // Высшее образование в России. 2023. Т. 32. № 4. С. 9–22. DOI: 10.31992/0869-3617-2023-32-4-9-22
6. Ладыжец Н.С. СОЦИАЛЬНЫЕ АСПЕКТЫ УПРАВЛЕНИЯ РИСКАМИ И ВОЗМОЖНОСТЯМИ ОПЕРЕЖАЮЩЕГО РАЗВИТИЯ НЕЙРОСЕТЕЙ // Вестник Удмуртского университета. Социология. Политология. Международные отношения. 2023. №2. URL: https://cyberleninka.ru/article/n/sotsialnye-aspekty-upravleniya-riskami-i-vozmozhnostyami-operezhayuschego-razvitiya-neyrosetey
7. Лёвин Б.А., Пискунов А.А., Поляков В.Ю., Савин А.В. Искусственный интеллект в инженерном образовании // Высшее образование в России. 2022. Т. 31. № 7. С. 79–95. DOI: 10.31992/0869-3617-2022-31-7-79-95
8. Малышев И.О., Смирнов А.А. ОБЗОР СОВРЕМЕННЫХ ГЕНЕРАТИВНЫХ НЕЙРОСЕТЕЙ: ОТЕЧЕСТВЕННАЯ И ЗАРУБЕЖНАЯ ПРАКТИКА // Международный журнал гуманитарных и естественных наук. 2024. №1-2 (88). URL: https://cyberleninka.ru/article/n/obzor-sovremennyh-generativnyh-neyrosetey-otechestvennaya-i-zarubezhnaya-praktika
9. Намиот Д.Е., Зубарева Е.В. О работе AI Red Team // International Journal of Open Information Technologies. 2023. №10. URL: https://cyberleninka.ru/article/n/o-rabote-ai-red-team
10. Оравек Дж.А. Последствия для академического мошенничества: расширение масштабов ответственного сотрудничества с людьми и ChatGPT // Журнал Interactive Учебные исследования. 2023. Vвыпуск 34. № 2. С. 213–237. URL: https://philarchive.org/archive/
11. Сазонов А.П. ИСПОЛЬЗОВАНИЕ ИИ В ПРОГРАММИРОВАНИИ // Universum: технические науки. 2024. №3 (120). URL: https://cyberleninka.ru/article/n/ispolzovanie-ii-v-programmirovanii
12. Сулейманова Д.О., Магомаев Т.Р. РОЛЬ CHATGPT В НАУКЕ О ДАННЫХ // Общество, экономика, управление. 2023. №2. URL: https://cyberleninka.ru/article/n/rol-chatgpt-v-nauke-o-dannyh-1
13. Alkaissi H, McFarlane S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing // Cureus. 2023. Vol. 15. № 2: e35179. DOI: 10.7759/cureus.35179
14. Nils Kobis, Luca D. Mossink Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry,Computers in Human Behavior, Volume 114, 2021, 106553, ISSN 0747-5632 — https://doi.org/10.1016/j.chb.2020.106553.
15. Hanna R. How and why ChatGPT failed The Turing Test. Against Professional Philosophy Website. Retrieved February 05, 2023 from: https://againstprofphil.org/2023/01/15/how-and-why-chatgpt-failed-the-turing-test/

Журнал «Современная наука: актуальные проблемы теории и практики»