УДК 004.912: 004.822
DOI: 10.17586/0021-3454-2024-67-11-958-968

АНАЛИЗ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ИСКУССТВЕННО СГЕНЕРИРОВАННЫХ ТЕКСТОВ

С. В. Кулешов, А. А. Зайцева*, А. Ю. Аксенов

Санкт-Петербургский федеральный исследовательский центр Российской академии наук,
Санкт-Петербург, Россия,
** cher@iias.spb.su*

Аннотация. Рассматривается новый тренд — формирование контента с применением инструментов и технологий искусственного интеллекта. Активное внедрение технологий искусственного интеллекта для генерации данных приводит к увеличению доли искусственно сгенерированных данных, которые необходимо выявлять в автоматическом режиме для предотвращения ошибок (недостоверности, введения в заблуждение). Предложены подходы к идентификации текстовых данных, созданных при помощи нейросетевых технологий, включающие эвристические правила, основанные на критерии зависимости объема реферата от порога реферирования, что позволяет проводить автоматическую оценку текстовых документов в мониторинговых и поисковых системах при обработке больших объемов неструктурированных данных. Полученные результаты закладывают технологическую базу для реализации широкого спектра практических решений по обеспечению интеллектуальной поддержки коллективного поведения участников в человекомашинных сообществах за счет разработки теоретических и технологических основ обработки неструктурированных данных.

Ключевые слова: *интернет-документы, искусственные нейронные сети, большая языковая модель, интернет-ресурсы, методы искусственного интеллекта, генерация данных*

Благодарности: работа выполнена при поддержке гос. заданием на 2024 г. № FFZF-2022-0005.

Ссылка для цитирования: *Кулешов С. В., Зайцева А. А., Аксенов А. Ю.* Анализ статистических характеристик искусственно сгенерированных текстов // Изв. вузов. Приборостроение. 2024. Т. 67, № 11. С. 958–968. DOI: 10.17586/0021-3454-2024-67-11-958-968.

ANALYSIS OF STATISTICAL CHARACTERISTICS OF ARTIFICIALLY GENERATED TEXTS

S. V. Kuleshov, A. A. Zaytseva*, A. Yu. Aksenov

St. Petersburg Federal Research Center of the RAS, St. Petersburg, Russia
** cher@iias.spb.su*

Abstract. A new trend is considered, namely, the formation of content using artificial intelligence tools and technologies. Active implementation of artificial intelligence technologies for data generation leads to an increase in the share of artificially generated data that must be identified automatically to prevent errors (unreliability, misleading). Approaches to identifying text data created using neural network technologies are proposed, including heuristic rules based on the criterion of dependence of the abstract volume on the abstracting threshold, which allows for automatic evaluation of text documents in monitoring and search systems when processing large volumes of unstructured data. The obtained results lay the technological basis for the implementation of a wide range of practical solutions to ensure intellectual support for the collective behavior of participants in human-machine communities through the development of theoretical and technological foundations for processing unstructured data.

Keywords: internet documents, artificial neural networks, large language model, Internet resources, artificial intelligence methods, data generation

Acknowledgments: the work was carried out with the support of the State assignment for 2024 No. FFZF-2022-0005.

For citation: Kuleshov S. V., Zaytseva A. A., Aksenov A. Yu. Analysis of statistical characteristics of artificially generated texts . *Journal of Instrument Engineering*. 2024. Vol. 67, N 11. P. 958–968 (in Russian). DOI: 10.17586/0021-3454-2024-67-11-958-968.

Введение. Активное внедрение технологий искусственного интеллекта во все области деятельности человека, включая интернет-пространство, привело к появлению нейросетевых инструментов генерации данных. Будем называть искусственно сгенерированными данными (ИСД) такие данные, которые сгенерированы с использованием искусственных нейронных сетей или алгоритмически.

На текущий момент тексты, сгенерированные подобным образом, используются в рекламных целях, при имитации активности сообщества, генерации новостей, написании учебных работ и в ряде других случаев.

В связи с тем, что искусственные тексты могут содержать фактические ошибки, не гарантируют достоверность и могут вводить в заблуждение, актуальной проблемой является изучение свойств ИСД и их идентификация (выявление).

Многие крупные компании, специализирующиеся на обработке данных, расценивают искусственно-сгенерированные данные (или искусственно-сгенерированный контент — AI-generated content) как особый вид контента, а также отмечают необходимость их идентификации и маркировки. Эту идею активно подхватили в СМИ и социальных сетях [1]. Вирусное распространение нейросетевых технологий создания ИСД (AI-generated content) и потребность автоматического или хотя бы автоматизированного их распознавания обусловили необходимость адаптации и расширения возможностей существующих технологий обработки неструктурированных данных.

Современное состояние исследований. В современной научной литературе под термином „AI-generated content“ понимается результат (контент), созданный с помощью больших языковых моделей (обученных на большом количестве данных) на основании пользовательских запросов (подсказок) [2].

В научных исследованиях, посвященных AI-generated content или ИСД, рассматривается широкий круг вопросов, начиная от ответственного подхода к формированию ИСД [3] и к информированию об ИСД [4] до создания ИСД с использованием федеративного обучения [5] и обнаружения ИСД в любой форме, включая изображения лиц [6] или иных объектов [7], а также текстов [8] и музыки [9].

В [10] представлен всесторонний обзор ИСД, в том числе проблемных аспектов, применительно к таким направлениям, как:

— применение ИСД в критически важных областях, где требуется высокая надежность и точность результатов (например, здравоохранение);

— поиск баланса между специализацией и обобщением в обучающих наборах данных для создания ИСД;

— непрерывное обучение и переобучение моделей для создания ИСД;

— улучшение способностей моделей к логическому выводу;

— масштабирование нейросетевых моделей;

— социальные проблемы, такие как предвзятость и этика.

В [11] представлен сравнительный анализ не только преимуществ ИСД, таких как эффективность и масштабируемость, „преодоление писательского блока“, но и недостатков, включая усугубление социального дисбаланса, негативное воздействие на процесс обучения, снижение креативности.

Обнаружение ИСД (AI-generated content detection) является актуальной проблемой, которой посвящено не одно исследование по данной теме [12]. Так, в [13] проанализирован метод обнаружения ИСД, использующий водяные знаки, и установлено, что злоумышленник может добавить в ИСД во избежание их обнаружения небольшое, незаметное для человека, изменение с водяным знаком. Как показали результаты данного исследования, обнаружение ИСД на основе использования водяных знаков не так устойчиво, как считалось ранее: простого расширения стандартных методов обнаружения ИСД до методов, использующих водяные знаки, недостаточно, поскольку в них не учитываются уникальные характеристики водяных знаков. В [14] были протестированы различные инструменты для обнаружения ИСД: результаты тестов

продемонстрировали необходимость улучшения таких инструментов, так как они отстают от развития генеративных сетей. Показано, что инструменты обнаружения ИСД иногда бывают полезны, но их нестабильная производительность и зависимость от сложности моделей искусственного интелекта (ИИ) требуют более комплексного подхода, включая ручную предобработку. Оригинальный метод обнаружения ИСД с помощью наиболее популярного генеративного чат-бота — ChatGPT — был рассмотрен в [15]. По итогам исследования установлено, что в последней версии данного чат-бота (ChatGPT 4) допущено большое количество ложноположительных результатов — более 95 % текстов, написанных человеком, распознано как ИСД. В [16] приведен вывод, что несмотря на важность технических решений для обнаружения ИСД, в настоящее время они недостаточно эффективны и должны быть дополнены этическими рекомендациями по использованию ИСД в авторских работах. В [17] представлен „метод обнаружения на уровне предложений“ (SeqXGPT), при котором обрабатываются отдельные предложения, но в контексте всего документа с учетом взаимосвязи между предложениями и общим контекстом. Данный метод достаточно успешен в решении задачи распознавания ИСД. В [18] на примере студенческих работ отмечается, что обнаружение ИСД не должно происходить в автоматическом режиме, а инструменты для обнаружения ИСД могут использоваться лишь в незначительной степени для поддержки принятия решений при подозрениях в нечестности студентов.

В связи с этим требуется дополнительное исследование возможностей автоматического выявления такого типа контента и алгоритмов идентификации ИСД, чему и посвящена настоящая статья.

Исследование свойств текстовых данных, сгенерированных с помощью нейросетевых технологий. Исторически первыми методами нейросетевой искусственной генерации данных были рекуррентные сети (RNN, LSTM), к которым впоследствии добавлен механизм внимания — attention [19–21]. Дальнейшим развитием стали искусственные нейронные сети (ИНС) GPT и BERT, на основе которых возникло большое количество подобных моделей. GPT (Generative Pre-Trained Transformer) — нейросетевая языковая модель, основанная на архитектуре „трансформер“ и парадигме самообученная (self-supervised learning) на большом корпусе текстовых данных и предназначенная для генерации (или продолжения) текста. ИНС GPT выполняет языковое моделирование, т. е. предсказание следующего слова (для некоторых языков — фрагмента слова) с учетом предыдущего контекста.

Фактически генерация текстов является результатом извлечения внутренних знаний языковой модели через определение левого контекста: в данном случае начальных токенов (синтактических элементов) фразы — промта. На практике это позволяет решать множество задач: отвечать на вопросы, суммаризировать (сжимать) текст и строить диалоговые системы. Подбор модификаций начальной фразы известен как „Prompt Engineering“.

Большие языковые модели фактически сохраняют слепок данных на какой-то определенный момент времени, что позволяет выявлять ИСД по артефактам, привязанным ко времени. Например, нейросетевая модель GPT-3.5 от OpenAI на вопрос „Какой сейчас год?“ дает ответ „Сейчас год 2022“.

Основные этапы генерации текста с помощью GPT [19–21]:

— входной текст преобразуется в последовательность идентификаторов-токенов (токенизируется);

— список токенов проходит через слой Embedding Layer ИНС и преобразуется в список эмбеддингов;

— к каждому эмбеддингу добавляется positional embedding, кодирующий положение токена в последовательности;

— список эмбеддингов трансформируется, проходя через несколько одинаковых блоков Transformer Decoder (этот этап проилюстрирован на рис. 1, *а*);

— после того как список эмбеддингов пройдет через последний блок последовательности трансформеров, эмбеддинг, соответствующий последнему токену, матрично умножается на все

тот же входной, но уже транспонированный слой Embedding Layer, и после применения функции SoftMax формируется распределение вероятностей следующего токена;

— из полученного на предыдущем шаге распределения выбирается следующий токен;

— полученный новый токен добавляется к входному тексту (рис. 1, *б*), предыдущие шаги повторяются необходимое число раз.

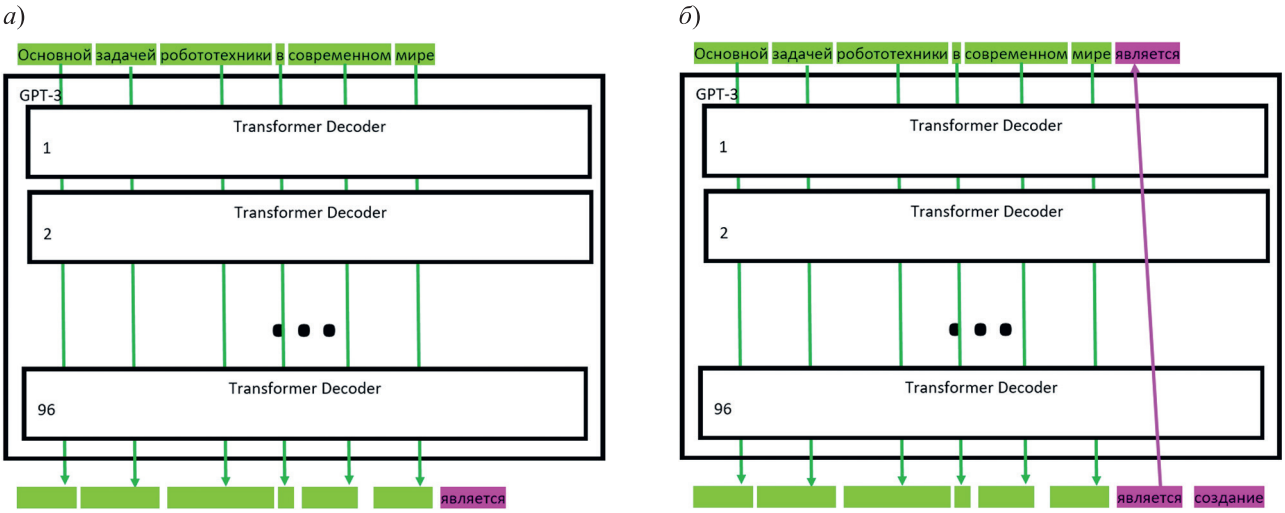

Рис. 1

Вариативность при генерации текста языковой моделью основана на распределении вероятностей следующего токена. Наиболее простой подход заключается в выборе токена с наибольшей вероятностью (greedy search). Основной недостаток такого подхода — опасность зацикливания генератора в локальных минимумах, что дает повторяющиеся фрагменты, например „The the the the...“.

Для формирования текста, обладающего свойствами естественного языка, используются следующие приемы:

— формирование образцов методом лучевого поиска (сэмплирование beam search);

— сэмплирование с температурой (параметр в распределении Больцмана) [22];

— сэмплирование с ограничением маловероятных токенов (nucleus sampling);

— методы файнтюнинга (донастройки модели для генерации в заданном стиле).

На текущий момент существует четыре поколения сетей GPT:

GPT-1 — первая версия модели состояла из 12 слоев и была обучена на 7000 книг. Как языковая модель она работала не очень корректно (особенно при генерации длинных текстов). Максимальный размер контекста у GPT-1 — 512 токенов.

GPT-2 — содержит 48 слоев и порядка 1,5 млрд параметров. Модель обучена на большем объеме данных, чем GPT-1, — к книгам добавлено 8 млн сайтов, что суммарно составляет 40 Гб текста. Архитектурно модель отличается незначительно, основные изменения касаются слоев нормализации. При этом GPT-2 оказалась способна генерировать более длинные связные тексты и решать при помощи prompt engineering множество новых классов задач. Максимальный размер контекста у GPT-2 — 1024 токенов.

GPT-3 — в этой версии были увеличены размер модели (в 10 раз, стало около 175 млрд параметров) и размер обучающей выборки (около 570 Гб текста). К архитектурным изменениям относится оптимизация параметра внимания. Данное обновление качественно улучшило результат генерации текстов, модель приобрела способность генерировать программный код (проект CODEX [23]). Максимальный размер контекста у GPT-3 — 2048 токенов.

GPT-4 — мультимодальная большая языковая модель 4-й версии. В техническом отчете GPT-4 параметры модели явно не приводятся, что обусловлено ссылкой на „конкурентную среду и последствия для безопасности крупномасштабных моделей“ [24]. Указывается, что модель может анализировать или генерировать тексты размером до 25 000 слов.

Известны системы, детектирующие искусственно сгенерированные данные:

— сервисы Pr-cy, Copyleaks для текстов с отдельной функцией проверки работ из образовательной сферы; сервис Gptzero, ориентированный на проверку научных текстов, а также Writer, Contentatscale и Originality;

— сервисы Ai or not и Maybe's AI Art Detector для изображений.

Нейросетевые технологии также могут использоваться для проверки создания контента на признак ИСД. Так, в ChatGPT можно загрузить текст размером от 1000 знаков для проведения такого анализа. Еще один детектор такого типа — Crossplag.

На практике эти инструменты оказываются не всегда эффективны. Если человек приложил некоторые усилия для улучшения ИСД, простые методы могут не распознать факт искусственной генерации контента. Так, текст можно доработать вручную или с использованием инструментов ИИ изменить его тональность, добавить детали и эмоции, изображение можно скорректировать в графических редакторах, применить фильтры. В этом случае перечисленные сервисы могут идентифицировать контент как уникальный.

Проведенный анализ позволил выделить следующие признаки искусственно сгенерированных данных (текстовых или графических) [8, 25–27]:

— в тексте, сгенерированном ИИ, очень редко встречаются опечатки и орфографические ошибки, не встречаются разные типы тире (-, – и —) или кавычек ("" и «»); в текстах, написанных человеком, такая идеальная точность встречается редко; соответственно отсутствие подобных ошибок можно использовать в качестве одного из признаков ИСД;

— текст состоит из общих фраз, в нем нет конкретики, деталей, а если встречаются названия, то не в сокращенной и привычной форме, а полностью: например, будет написано не „МГУ“, а „Московский государственный университет им. М. В. Ломоносова“; на текущей стадии развития генераторы формируют ИСД, не содержащие экспрессивных и эмоционально окрашенных деталей;

— системы ИИ при генерации контента выдают дублирующие элементы: повторение слов и фраз или одинаковые детали на картинке; например, генератор изображений „Шедеврум“ (разработчик „Яндекс“) по запросу „Нарисуй руку с пятью пальцами, не с четырьмя и не с шестью“ генерирует изображение руки с шестью пальцами;

— при увеличении масштаба сгенерированных изображений заметны нестыковки вроде отсутствия теней, разрыва линий, резких перепадов цвета без градиента; чаще всего искажения и артефакты встречаются в фоне (например, линия горизонта зигзагами); при создании изображения нейросети не учитывают законы физики и геометрии;

— при сгенерированных изображениях, содержащих лица, создается сюрреалистический фон для портрета, размазываются участки волос, появляется значительная асимметрия в парных деталях (глаза, уши), появляются „шумы“ в виде горизонтальных или вертикальных полос на однотонных участках или градиент в зонах резкой смены цвета, например между шеей и воротником;

— созданный человеком документ может иметь изменения стиля и тона по всему тексту, тогда как содержание документа, формируемого системой ИИ, остается одинаковым.

Идентификация текстовых данных, сгенерированных с помощью нейросетевых технологий. Существующие подходы к идентификации ИСД базируются на использовании особенностей GPT-подобных моделей на основе нейронных сетей, наиболее известной является ChatGPT.

ChatGPT основана на архитектуре GPT, которая позволяет модели обрабатывать длинные последовательности текста и улавливать контекстуальные зависимости между словами. Благодаря этой архитектуре ChatGPT может генерировать тексты, которые кажутся естественными и связными.

Одним из ключевых элементов ChatGPT является механизм внимания, позволяющий модели фокусироваться на семантически важных фрагментах входного текста и использовать эту информацию для генерации ответа. Механизм внимания также помогает модели учитывать большее количество элементов контекста и создавать более качественные и связные ответы.

Процесс обучения модели ChatGPT состоит из двух этапов: предварительного обучения и дообучения. На первом этапе модель обучается на больших объемах неразмеченных текстов, чтобы она научилась улавливать общую структуру и грамматические правила естественного языка. На втором этапе модель обучается на задаче генерации текстовых ответов на вопросы. Этот этап позволяет модели становиться более специализированной и настраивать свои ответы, соответствуя заданной задаче.

Однако при генерации текстов с использованием ChatGPT-подобных инструментов могут возникать определенные проблемы, такие как непоследовательность ответов, недостаточная информативность, отклонение от темы и возможные проблемы со стилистикой. Эти проблемы связаны с тем, что модель может придумывать несуществующие факты, избегать прямых ответов или давать неадекватные комментарии. Именно эти особенности используются при разработке методов идентификации ИСД, однако если проведены соответствующие фильтрация или контроль, идентификация таких ИСД становится более сложной.

Известны следующие подходы к идентификации ИСД:

— обнаружение non-contextual tokens (неконтекстуальных токенов): ChatGPT-модель часто вставляет специальные токены или маркеры в свои ответы для указания начала или конца ответа; поиск таких токенов в сгенерированном тексте позволяет делать выводы об использовании модели;

— анализ статистических свойств сгенерированного текста, таких как распределение слов, грамматика или стилистические особенности, может включать проверку использования редких или специфичных слов и фраз, обнаружение неправильной грамматики или плохого стиля;

— методы метаданных: для обнаружения применения ChatGPT-модели используется дополнительная информация, такая как сведения о запросах, метаданные сеанса и временные отметки;

— анализ контекста вопросов и ответов: если в ответах модели отсутствуют устойчивые или явные связи с предыдущими вопросами либо контекст быстро теряется, это может быть признаком автоматической генерации.

Среди прочих методов проверки также следует отметить метод сравнения следующих токенов при совпадающих предыдущих (идентичном левом контексте) в проверяемом тексте и в тексте, генерируемом GPT-подобной моделью; проверка образцов путем их поиска в интернет-архивах.

Подход к идентификации ИСД на основе эвристических правил. Опыт использования разработанного авторами ранее метода [28] идентификации текстов, искусственно сгенерированных алгоритмически, дает предпосылку к его адаптации для применения в текстах, сгенерированных с помощью нейросетевых GPT-моделей.

Идея метода [28] основана на оценке скорости уменьшения объема автоматически сформированного реферата текста на каждом шаге реферирования (при последовательном увеличении порога ε). Приведем краткое описание метода.

Пусть $s \in T$, где s — предложения, являющиеся элементами множества предложений текста T. В этом случае рефератом текста называется множество F, если $F \subset T$, $|F| < |T|$.

Рефератом является множество F_ε на каждом шаге $\varepsilon = 1, 2, \ldots, n$, которое формируется из предложений s исходного текста T по правилам $s \in F_\varepsilon$, если $\rho(s) \geq \varepsilon$, где $\rho(s)$ — рейтинг предложения; значение n определяется условием $|F_n| = 0$.

Рейтингом предложения считается максимальный рейтинг элементов множества K_s двуграмм, отображающих синтаксическую связь k между двумя словами, входящими в предложение s, рассчитываемый по формуле

$$\rho(s) = \max_{k \in K_s} |L_k|, \; s \in L_k, \tag{1}$$

где L_k — множество предложений, содержащих синтаксическую связь k.

В качестве критериев принадлежности текста к алгоритмически сгенерированному в работе [28] были использованы следующие правила (П1.1–П1.3):

П1.1: число различных значений рейтинга на последовательности шагов меньше 3 или первое значение меньше 20.

П1.2: число различных значений меньше 5.

П1.3: число первых ненулевых значений меньше 4 или число подряд идущих одинаковых значений больше 4.

Критерии принадлежности текста к ИСД, полученным с использованием нейросетевых технологий, предлагается сформировать на основе критериев принадлежности текста к алгоритмически сгенерированному.

Для этого были сформированы наборы ИСД на основе большой языковой модели mistral-7b-instruct-v0.1.Q4_0 [29], на которую в качестве промтов (левого контекста) подавались текстовые строки согласно таблице. В качестве значений <title> в промтах использовались заголовки финансовых новостей за период с 2021 по 2023 гг., названия научных статей (из индекса научного цитирования), а также заголовки художественных произведений на русском языке. Результатом работы модели является текст на русском или английском языке, который помещается в набор.

Тестовый набор	Набор заголовков	Промт	Объем выборки (русский/английский)
A	Названия научных журналов	Напиши рекламный текст про <title> на русском языке	83/26
B	Заголовки новостей на финансовую тематику	Напиши новость <title> на русском языке	83/23
C	Заголовки художественных произведений	Напиши сказку <title> на русском языке	135/68

Схема формирования наборов ИСД приведена на рис. 2.

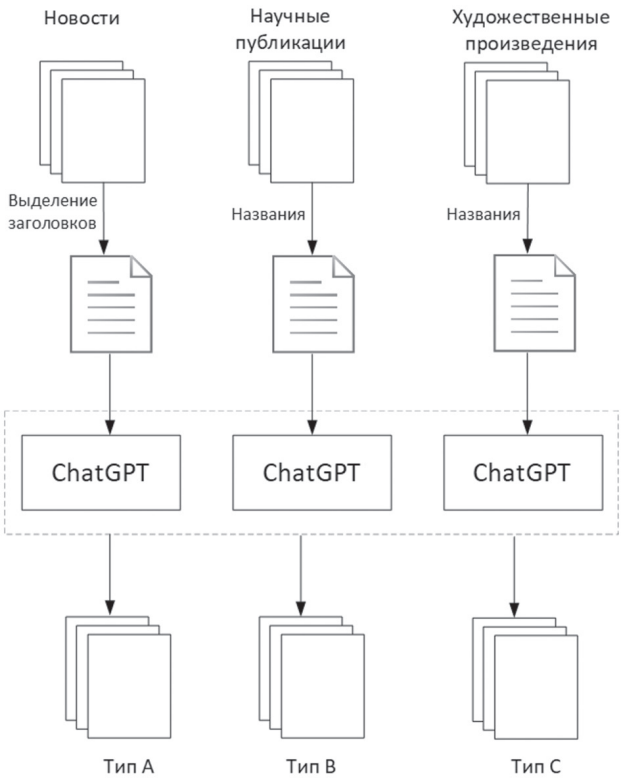

Рис. 2

На рис. 3, *а*, *б*, *в* приведен вид кривых зависимости объема реферата (S) от порога реферирования (ε) для ИСД — наборы A, B, C соответственно.

Согласно проведенным экспериментам статистические характеристики сгенерированных ChatGPT-моделью текстов, выявляемые на основе ассоциативно-онтологического (графового) представления текста, аналогичны характеристикам текста того же вида, созданного человеком. Основным фиксируемым отличием кривых зависимости $S(\varepsilon)$ для ИСД от кривых для текстов, созданных человеком, является быстрое уменьшение объема реферата S до нуля после некоторого значения ε (см. рис. 3).

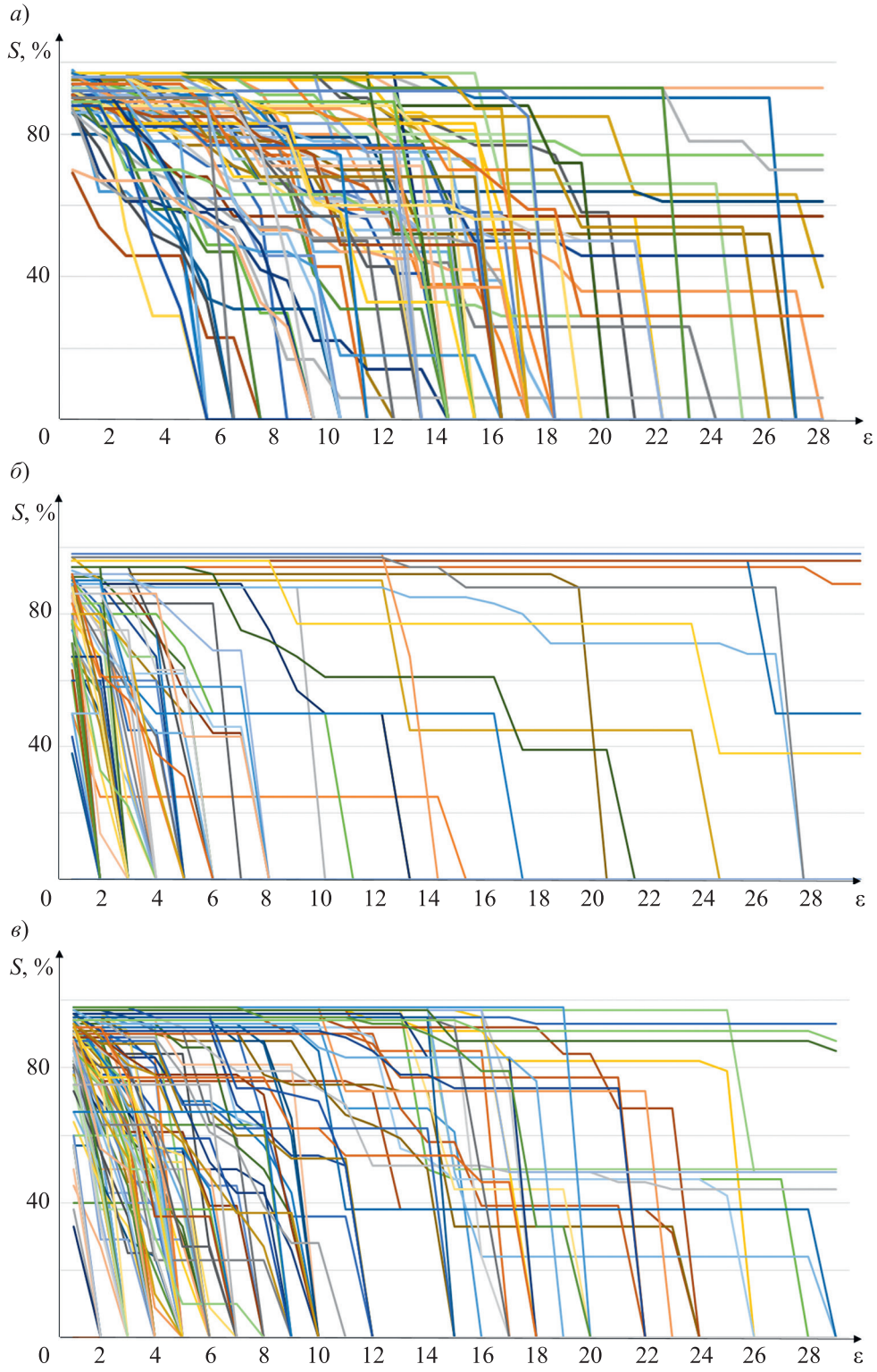

Рис. 3

На основании проведенных экспериментов, с учетом [28], сформулированы эвристические правила идентификации ИСД, полученных с использованием нейросетевых технологий, на основе критерия зависимости объема реферата от порога реферирования (П2.1–П2.3):

П2.1: число различных значений меньше 10.

П2.2: последнее ненулевое значение больше 40.

П2.3: число первых ненулевых значений меньше 4 или число подряд идущих одинаковых значений больше 4.

Заключение. Рассмотрены особенности работы и применения нейросетевых инструментов генерации данных.

Выявлено значительное увеличение доли искусственно сгенерированных с применением нейросетевых подходов данных в текстовом виде, которые можно рассматривать как один из видов слабоструктурированных данных. Введено понятие искусственно сгенерированных данных, дано его определение и исследованы свойства ИСД. Разработаны технологии идентификации текстовых данных, сгенерированных с помощью нейросетевых технологий, содержащие эвристические правила на основе критерия зависимости объема реферата (полученного с использованием ассоциативно-онтологического подхода) от порога реферирования, что позволяет автоматически оценивать качество текстовых документов в мониторинговых и поисковых системах при обработке больших объемов неструктурированных данных.

Предметом дальнейших исследований является верификация сформулированных эвристических правил на других нейросетевых моделях и их алгоритмическая реализация.

Все приведенные в статье фактические числовые и экспериментальные данные актуальны на начало 2024 г.

СПИСОК ЛИТЕРАТУРЫ

1. YouTube обяжет маркировать контент, созданный нейросетями [Электронный ресурс]: https://www.fontanka.ru/2023/11/14/72913286/, 27.06.2024.
2. *Fang X., Che Sh., Mao M., Zhang H., Zhao M., Zhao X.* Bias of AI-Generated Content: An Examination of News Produced by Large Language Models [Электронный ресурс]: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4574226, 27.06.2024.
3. *Chen Ch., Fu J., Lyu L.* A Pathway Towards Responsible AI Generated Content. 2023. DOI: 10.48550/arXiv.2303.01325.
4. *Wahle J.Ph., Ruas T., Mohammad S.M., Meuschke N., Gipp B*. AI Usage Cards: Responsibly Reporting AI-Generated Content // Proc. of ACM/IEEE Joint Conf. on Digital Libraries (JCDL 2023), June 2023, Mexico, Santa Fe. 2023. P. 282–284.
5. *Huang X., Li P., Du H., Kang J., Niyato D., Kim D.I., Wu Y*. Federated Learning-Empowered AI-Generated Content in Wireless Networks. 2023. DOI: 10.48550/arXiv.2307.07146.
6. *Gragnaniello D., Marra F., Verdoliva L.* Detection of AI-Generated Synthetic Faces. Handbook of Digital Face Manipulation and Detection // Advances in Computer Vision and Pattern Recognition. 2022. P. 191–212.
7. *Xi Z., Wenmin H., Kangkang W., Weiqi L., Peijia Zh.* AI-Generated Image Detection using a Cross-Attention Enhanced Dual-Stream Network // Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Nov. 2023, Taiwan, Taipei. P. 1463–1470.
8. *Weber-Wulff D., Anohina-Naumeca A., Bjelobaba S., Foltýnek T., Guerrero-Dib J., Popoola O., Šigut P., Waddington L.* Testing of Detection Tools for AI-Generated Text. 2023. DOI: 10.48550/arXiv.2306.15666.
9. *Joo-Wha H., Fischer K., Ha Y., Zeng Y*. Human, I wrote a song for you: An experiment testing the influence of machines' attributes on the AI-composed music evaluation//Computers in Human Behavior. 2022. Vol. 131. 107239.
10. *Cao Y. Li S., Liu Y., Yan Zh., Dai Y., Yu Ph., Sun L*. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. 2023. DOI: 10.48550/arXiv.2303.04226.
11. *Wu J., Wensheng G., Zefeng Ch., Shicheng W., Hong L*. AI-Generated Content (AIGC): A Survey. 2023. DOI: 10.48550/arXiv.2304.06632.
12. *Ruchika L., Priyanka Bh., Neha V., Anshika J*. AI-Generated Text Detection: A Review // Intern. Journal of Creative Research Thoughts (IJCRT). 2023. Vol. 11(10). P. d784–d789.
13. *Zhengyuan J., Jinghuai Zh., Neil Zh.G*. Evading Watermark based Detection of AI-Generated Content // Proc. of the ACM SIGSAC Conf. on Computer and Communications Security (CCS '23), Nov. 2023, Copenhagen. 2023. P. 1168–1181.

14. *Elkhatat A., Elsaid Kh., Almeer S.* Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text // Intern. Journal for Educational Integrity. 2023. Vol. 19. P. 17.
15. *Elkhatat A. M.* Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities // Intern. Journal for Educational Integrity. 2023. Vol. 19. P. 15. DOI: 10.1007/s40979-023-00137-0.
16. *Otterbacher J.* Why technical solutions for detecting AI-generated content in research and education are insufficient// Patterns. 2023. Vol. 4(7). P. 100796.
17. *Pengyu W., Linyang K. R., Botian J., Dong Zh., Xipeng Q.* SeqXGPT: Sentence-Level AI-Generated Text Detection // Proc. of the Conf. on Empirical Methods in Natural Language Processin, Dec. 2023. Singapore. 2023. P. 1144–1156.
18. *Price G. Sakellarios M.* The Effectiveness of Free Software for Detecting AI-Generated Writing // Intern. Journal of Teaching, Learning and Education. 2023. Vol. 2. P. 31–38.
19. *Qu Y., Liu P., Song W., Liu L., Cheng M.* A Text Generation and Prediction System: Pre-training on New Corpora Using BERT and GPT-2 // IEEE 10th Int. Conf. on Electronics Information and Emergency Communication (ICEIEC), July 2020, China, Beijing. 2020. P. 323–326.
20. *Chen W., Su Y., Yan X., Wang W. Y.* KGPT: Knowledge-Grounded Pre-Training for Data-to-Text Generation. [Электронный ресурс]: https://arxiv.org/abs/2010.02307, 27.06.2024.
21. GPT для чайников: от токенизации до файнтюнинга [Электронный ресурс]: https://habr.com/ru/articles/599673/, 27.06.2024.
22. *Ackley D., Hinton G., Sejnowski T.* A learning algorithm for Boltzman nmachines//Cognitive Science. 1985. Vol. 9. N 1. P. 147–169.
23. OpenAI Codex [Электронный ресурс]: https://openai.com/blog/openai-codex, 27.06.2024.
24. GPT-4 Technical Report. OpenAI [Электронный ресурс]: https://cdn.openai.com/papers/gpt-4.pdf, 27.06.2024.
25. GPTZero [Электронный ресурс]: https://gptzero.me/technology, 27.06.2024.
26. *Chaka C.* Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools//Journal of Applied Learning and Teaching. 2023. Vol. 6(2). DOI: 10.37074/jalt.2023.6.2.12.
27. *Yang X., Cheng W., Petzold L., Wang W.Y., Chen H.* DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text//ArXiv, abs/2305.17359. 2024.
28. *Кулешов С. В., Зайцева А. А., Марков С. В.* Ассоциативно-онтологический подход к обработке текстов на естественном языке // Интеллектуальные технологии на транспорте. 2015. № 4. С. 40–45.
29. *Jiang A. Q.* et al. Mistral 7B [Электронный ресурс]: https://arxiv.org/abs/2310.06825, 27.06.2020.

СВЕДЕНИЯ ОБ АВТОРАХ

Сергей Викторович Кулешов — д-р техн. наук, профессор РАН; СПбФИЦ РАН, СПИИРАН, лаборатория автоматизации научных исследований; гл. научный сотрудник; E-mail: kuleshov@iias.spb.su

Александра Алексеевна Зайцева — канд. техн. наук; СПбФИЦ РАН, СПИИРАН, лаборатория автоматизации научных исследований; ст. научный сотрудник; E-mail: cher@iias.spb.su

Алексей Юрьевич Аксенов — канд. техн. наук; СПбФИЦ РАН, СПИИРАН, лаборатория автоматизации научных исследований; ст. научный сотрудник; E-mail: a_aksenov@iias.spb.su

Поступила в редакцию 23.07.24; одобрена после рецензирования 01.08.24; принята к публикации 23.09.24.

REFERENCES

1. https://www.fontanka.ru/2023/11/14/72913286/. (in Russ.)
2. Fang X., Che Sh., Mao M., Zhang H., Zhao M., Zhao X. *Sci. Rep.*, 2024, no. 1(14), pp. 5224, doi: 10.1038/s41598-024-55686-2.
3. Chen Ch., Fu J., Lyu L. *arXiv:2303.01325v3*, 27 Dec. 2023, https://doi.org/10.48550/arXiv.2303.01325.
4. Wahle J.Ph., Ruas T., Mohammad S.M., Meuschke N., Gipp B. *Proc. of 2023 ACM/IEEE Joint Conf. on Digital Libraries* (JCDL 2023), Mexico, Santa Fe, June 2023, pp. 282–284.
5. https://doi.org/10.48550/arXiv.2307.07146.
6. Gragnaniello D., Marra F., Verdoliva L. *Advances in Computer Vision and Pattern Recognition*, 2022, pp. 191–212.
7. Xi Z., Wenmin H., Kangkang W., Weiqi L., Peijia Zh. *Proc. of 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference* (APSIPA ASC), Taiwan, Taipei, November 2023, pp. 1463–1470.
8. https://doi.org/10.48550/arXiv.2306.15666.
9. Joo-Wha H., Fischer K., Ha Y., Zeng Y. *Computers in Human Behavior*, 2022, vol. 131, art. no. 107239.

10. https://doi.org/10.48550/arXiv.2303.04226.
11. https://doi.org/10.48550/arXiv.2304.06632.
12. Ruchika L., Priyanka Bh., Neha V., Anshika J. *Intern. J. of Creative Research Thoughts* (IJCRT), 2023, no. 10(11), pp. d784–d789.
13. Zhengyuan J., Jinghuai Zh., Neil Zh.G. *Proc. of the 2023 ACM SIGSAC Conf. on Computer and Communications Security* (CCS '23), Denmark, Copenhagen, November 2023, pp. 1168–1181.
14. Elkhatat A., Elsaid Kh., Almeer S. *Intern. J. for Educational Integrity*, 2023, vol. 19, pp. 17.
15. Elkhatat A.M. *Intern. J. for Educational Integrity*, 2023, vol. 19, pp. 15, https://doi.org/10.1007/s40979-023-00137-0.
16. Otterbacher J. *Patterns*, 2023, no. 7(4), pp. 100796.
17. Pengyu W., Linyang K.R., Botian J., Dong Zh., Xipeng Q. *Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing 2023*, Singapore, December 2023, pp. 1144–1156.
18. Price G. Sakellarios M. *Intern. J. of Teaching, Learning and Education*, 2023, vol. 2, pp. 31–38.
19. Qu Y., Liu P., Song W., Liu L., Cheng M. *IEEE 10th Intern. Conf. on Electronics Information and Emergency Communication* (ICEIEC), China, Beijing, July 2020, pp. 323–326.
20. https://arxiv.org/abs/2010.02307.
21. https://habr.com/ru/articles/599673/. (in Russ.)
22. Ackley D., Hinton G., Sejnowski T. *Cognitive Science*, 1985, no. 1(9), pp. 147–169.
23. *OpenAI Codex*, https://openai.com/blog/openai-codex.
24. *GPT-4 Technical Report*. OpenAI, https://cdn.openai.com/papers/gpt-4.pdf.
25. *GPTZero*, https://gptzero.me/technology.
26. Chaka C. *Journal of Applied Learning and Teaching*, 2023, no. 2(6), https://doi.org/10.37074/jalt.2023.6.2.12.
27. Yang X., Cheng W., Petzold L., Wang W.Y., Chen H. *ArXiv, abs/2305.17359*, https://www.semanticscholar.org/paper/DNA-GPT%3A-Divergent-N-Gram-Analysis-for-Detection-of-Yang-Cheng/08145978da4c8912f4a05444a6bbf048778dc4af.
28. Kuleshov S.V., Zaytseva A.A., Markov S.V. *Intellectual Technologies on Transport*, 2015, no. 4, pp. 40–45. (in Russ.)
29. https://arxiv.org/abs/2310.06825.

DATA ON AUTHORS

Sergey V. Kuleshov — Dr. Sci., Professor; St. Petersburg Federal Research Center of the RAS, St. Petersburg Institute for Informatics and Automation of the RAS, Laboratory of Automation of Scientific Research, Chief Researcher; E-mail: kuleshov@iias.spb.su

Alexandra A. Zaytseva — PhD; St. Petersburg Federal Research Center of the RAS, St. Petersburg Institute for Informatics and Automation of the RAS, Laboratory of Automation of Scientific Research, Senior Researcher; E-mail: cher@iias.spb.su

Alexey Yu. Aksenov — PhD; St. Petersburg Federal Research Center of the RAS, St. Petersburg Institute for Informatics and Automation of the RAS, Laboratory of Automation of Scientific Research, Senior Researcher; E-mail: a_aksenov@iias.spb.su

Received 23.07.24; approved after reviewing 01.08.24; accepted for publication 23.09.24.