Чижов Дмитрий Александрович
студент 2 курса магистратуры факультета Международной журналистики
МГИМО МИД России
119454 Москва, проспект Вернадского, 76.
E-mail: dmitry4izhov@gmail.com

СТРАТЕГИИ ПРИМЕНЕНИЯ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В БОРЬБЕ С ИНФОРМАЦИОННЫМ ФАЛЬСИФИКАТОМ

Научная статья на основе итогов круглого стола «Фальсификация истории и средства массовой информации» в рамках проекта исторической памяти «РОССИЯ не заБУДЕТ» 25.10.2023 года, МГИМО

Аннотация: Идея использования передовых технологических инструментов для борьбы с информационным фальсификатом в наше время приобретает особую актуальность. Эта тенденция связана с тем, что новейшие цифровые технологии с каждым днем получают большее распространение и занимают все более значимое место в международных коммуникациях. В свою очередь, информационный фальсификат, дезинформация и фейки становятся серьезной угрозой, поскольку оказывают негативное влияние на общественное мнение, процессы принятия решений и политическую ситуацию в стране. В связи с этим представляется необходимым подробнее изучить возможности использования нейросетей в качестве действенного инструмента для своевременного реагирования на производство информационного фальсификата и противодействия его распространению в современном информационном поле.
Ключевые слова: нейросеть, искусственный интеллект, GPT-4, чат-бот, тиражируемость, алгоритмы, фейки, информационный фальсификат, дезинформация, фактчекинг, верификация, международные коммуникации.

Технологический прогресс в области искусственного интеллекта значительно увеличил тиражируемость фейковой информации и создал инструменты для ее автоматизированной генерации в глобальном масштабе. Растущие объемы производимого информационного фальсификата сегодня представляют угрозу для общества и серьезный вызов для современной науки. Неконтролируемо распространяемая дезинформация может оказывать негативное влияние на общественное мнение, процессы принятия решений и политическую ситуацию в стране. К наиболее опасным последствиям

распространения заведомо ложных сведений относятся: 1) репутационные риски, в том числе уменьшение доверия к СМИ; 2) манипулирование общественным мнением; 3) вмешательство в личную жизнь граждан; 4) организованная преступность и международный терроризм; 5) угроза социальной стабильности, национальной и международной безопасности. Специалисты сходятся во мнении, что особенно уязвимыми к распространению дезинформации оказываются сферы политики и здравоохранения, что было видно во время пандемии COVID-19 [3, 62-66]. Таким образом, сегодня необходимыми представляются решения, которые позволят системно противостоять дезинформации в стратегически важных для государства сферах, и в этом контексте нейросети имеют существенный, не до конца раскрытый потенциал и могут в перспективе стать действенным инструментом в борьбе против информационного фальсификата в международных коммуникациях.

Генеративная нейросеть представляет собой систему соединённых и взаимодействующих между собой искусственных нейронов. Это программа, которая позволяет автоматически получать информацию в осмысленном виде и способна выполнять поставленные перед ней творческие задачи. Сегодня определить сгенерированный искусственным интеллектом текст для человека становится все сложнее: нейросеть может сымитировать стиль человеческой речи, написать аналитический материал с использованием специфической терминологии, составить документ с соблюдением заданных правил. Благодаря имитации нейронной связи человеческого мозга, искусственный интеллект обладает значительной убедительностью, а также оперативностью, превосходящей человеческую.

Основные преимущества современных нейросетей в случае недобросовестного использования можно без труда обратить во вред. В этой связи внимание на себя обращает открытое письмо с призывом приостановить эксперименты с искусственным интеллектом, подписанное главами крупнейших мировых технологических корпораций и опубликованное на

сайте некоммерческой организации Future of Life Institute 22 марта 2023 года. В письме исследователи в сфере высоких технологий и искусственного интеллекта призвали ввести всемирный мораторий на полгода на обучение систем искусственного интеллекта, которые были бы мощнее мультимодальной языковой модели GPT-4 (Generative Pre-trained Transformer 4). Озабоченность научного сообщества вызвала способность генеративного искусственного интеллекта выполнять «человеческие» задачи. В письме особо подчёркивалась возможность нейросетей генерировать заведомо неверные сведения: «Мы должны задаться вопросом: следует ли нам позволить машинам заполнить наши информационные каналы пропагандой и ложью?» [6]. Мощные ИИ-системы должны разрабатываться и совершенствоваться лишь при условии, что человечество будет уверено, что их воздействие будет позитивным, а риски - управляемыми, для чего необходимо разработать и внедрить соответствующие протоколы безопасности, резюмировали авторы открытого письма. Документ был опубликован после отчета разработчиков компании OpenAI о возможностях GPT-4, из которого следовало, что новая языковая модель может быть использована для генерации убедительной дезинформации [5, 46]. В ходе ряда тестов в Alignment Research Center OpenAI продемонстрировала способность чат-бота запустить фишинговую атаку и скрыть все доказательства собственного мошеннического поведения. Коллективную озабоченность относительно этой темы отчетливо зафиксировала единогласно принятая Генеральной Ассамблеей ООН в марте этого года резолюция о регулировании искусственного интеллекта, в которой утверждается что ненадлежащее или злонамеренное проектирование, разработка, внедрение и использование систем искусственного интеллекта создает риски, способные подорвать целостность информации и доступ к ней, а также отрицательно сказаться на защите, продвижении и осуществлении прав человека [7, 3]. Все это доказывает небезосновательность растущего беспокойства вокруг неконтролируемого внедрения и применения нейросетей. Однако искусственный интеллект подходит далеко не только для

создания дезинформации, но и для обнаружения и борьбы с этим явлением. Поэтому особую значимость в современных реалиях приобретает анализ потенциальных преимуществ автоматизированного обнаружения дезинформации и успешной интеграции технологических решений на базе искусственного интеллекта в медиасферу.

Обобщая современные достижения использования искусственного интеллекта для борьбы с информационным фальсификатом, можно выделить несколько основных преимуществ нейросетей для выполнения этой задачи.

Мультимодальное обнаружение. Распространение дезинформации в интернете, вызванное процессом цифровизации, значительно активизировало работу по проверке фактов, публикуемых СМИ и тиражируемых пользователями Сети. В данном контексте под проверкой фактов понимается процесс верификации сообщаемой информации - подтверждения достоверности содержащихся утверждений путем подбора авторитетного источника и сверки с ним. Этот вид деятельности всегда оставался неотъемлемой частью журналистской практики, однако вместе с ростом объемов и разнообразия информационного фальсификата в медиасфере увеличилась и необходимость его расширения за счет возможностей искусственного интеллекта. Существующие достижения в этой области связаны с использованием технологий мультимодального автоматического обнаружения - алгоритмов, основанных на умении нейросетей считывать текстовые и визуальные сигналы обрабатываемых данных для выявления и категоризации дезинформации. Мультимодальное обнаружение позволяет вывести процесс верификации данных на полностью автоматизированный уровень и потому имеет потенциал в качестве инструмента противодействия информационному фальсификату.

Машинное обучение. В связи с ограниченностью времени и человеческих ресурсов, требуемых для выявления новых видов дезинформации, актуальность приобретают технологии непрерывного автоматизированного обучения искусственного интеллекта. Перспективным

шагом в этом направлении можно считать использование алгоритмов машинного обучения. Благодаря этой технологии искусственный интеллект может не просто взять на себя долю «рутинных» задач, но будет автономно совершенствовать собственные алгоритмы проверки на основе обрабатываемых данных и адаптировать их под выявление новых видов информационного фальсификата. Стоит отметить, что в Российской Федерации внедрением технологий искусственного сбора и верификации данных с использованием машинного обучения сегодня уже активно занимаются многие государственные ведомства и авторитетные медиа [2]. Это позволяет освободить сотрудников редакции СМИ для выполнения более трудоемких задач, требующих исключительно нестандартных решений, что в условиях конкуренции на медиарынке является несомненным преимуществом. Примечательно, что отечественные средства массовой информации зачастую используют для этого собственные цифровые разработки в области искусственного интеллекта. Собственное программное обеспечение вместо готовых компьютерных решений позволяет минимизировать возможные риски.

Анализ текста. Автоматизированная проверка информации может включать в себя анализ целого текста, выполняемый посредством соответствующих алгоритмов. Помимо семантического и синтаксического анализа, основанного на встречаемости слов и их связях, в этом случае также применяются проверка источников информации, перекрестная проверка данных, анализ объективности приводимой в тексте аргументации и т.д. Таким образом, задача текстового анализа может быть определена как извлечение значимой, полезной и ранее неизвестной информации из текстовых данных.

Обработка естественного языка. Высокую тиражируемость фейковым новостям обеспечивают социальные сети, мессенджеры и популярные развлекательные ресурсы. В системах машинного обучения представление данных существенно влияет на точность результатов, а контент, которым

делятся пользователи социальных сетей имеет, как правило, неструктурированную форму, что приводит к затруднению процесса верификации. Такие неструктурированные данные, извлекаемые нейросетями из социальных сетей, необходимо преобразовывать в структурированный формат с помощью методов интеллектуального анализа текста. Мониторинг социальных сетей на основе технологий обработки текстов на естественном языке (NPL) может сыграть решающую роль в снижении степени распространения дезинформации.

Говоря о существующих практических вызовах, стоит перечислить главные факторы, в той или иной мере снижающие эффективность современного применения искусственного интеллекта против распространения информационного фальсификата. Приводимые ниже проблематичные аспекты широкого использования нейросетей требуют первостепенного решения.

Прозрачность. Для эффективного внедрения и использования технологий верификации данных и автоматизированного реагирования на информационный фальсификат необходимо повысить общественное доверие к технологиям искусственного интеллекта в целом. Для этого рекомендуется делать акцент на расширении понимания принципов функционирования нейросетей среднестатистическим пользователем. Речь идет об объяснении, что в настоящий момент не разработано исключительно автономного универсального алгоритма, способного принимать решения полностью независимо от его оператора. Важно прийти к общему пониманию того, что существуют специализированные технологии искусственного интеллекта, большинство из которых имеет рутинное применение - как правило, они используются для фильтрации больших массивов данных, составления статистики, подсчета, маркировки и перевода информации. Поэтому представляется оправданным при возможности не скрывать факт применения нейросетей в рабочем процессе.

Этичность. Одна из наиболее распространенных проблем, связанных с использованием искусственного интеллекта и применением технологий машинного обучения для выявления дезинформации, связана с вопросом этических соображений: правильно ли человеку делегировать выполнение ответственных задач нейросети. Существующие в общественном сознании предубеждения относительного перспектив развития искусственного интеллекта создают препятствия для широкого применения этих технологий на практике. Метод решения этой проблемы граничит с описанным выше подходом к прозрачности применения технологий на базе ИИ, однако избавиться от активно тиражируемых произведениями массовой культуры мифов об искусственном интеллекте, который якобы неизбежно заменит человеческий, в краткосрочной перспективе представляется невозможным. Тем не менее, специалистам стоит начать активную работу в этом направлении и расширять понимание основных преимуществ и многогранности применения нейросетей. Постепенному пониманию важнейшего тезиса о том, что искусственный интеллект освободит человека для выполнения других задач, а не заменит его, способствует проведение просветительских мероприятий с участием ведущих авторитетных специалистов области.

Потребность в регулярном обновлении больших объемов данных. Одной из первых серьезных проблем, возникших на начальном этапе разработки нейросетей, стала нехватка регулярно обновляемой информации для обучения моделей. Несмотря на то, что неотъемлемым условием для совершенствования технологии автоматической классификации новостных текстов посредством искусственного интеллекта является создание обширных информационных хранилищ, обеспечение необходимого уровня качества данных обладает не меньшей значимостью для разработки эффективных алгоритмических решений по борьбе с дезинформацией. Специалисты рекомендуют полагаться на наборы данных, содержащие статьи, которые были предварительно помечены независимыми экспертами. В первую очередь, для увеличения точности, с которой нейросеть идентифицирует информационный

фальсификат, необходимо создание достаточно большой и разнообразной обучающей выборки фейковых новостей. Исследователи утверждают, что подобные базы данных должны состоять из образцов как ложной, так и правдивой информации со сбалансированным распределением по различным темам [1, 11]. В настоящий момент количество подобных наборов данных невелико, поскольку индивидуальная маркировка информационного контента требует много времени. Тем не менее, важным шагом для дальнейшей работы в этом направлении можно считать специализированные интернет-страницы с базами предварительно верифицированной информации. В качестве примеров одного из отечественных проектов такого рода можно назвать информационный ресурс объясняем.рф, запущенный правительством Российской Федерации весной 2022 года с целью информирования граждан по актуальным вопросам и недопущения распространения фейков [4].

Контекстное восприятие. Еще один вектор для дальнейшего совершенствования технологий искусственного интеллекта с целью противостояния информационному фальсификату в международных коммуникациях задает необходимость учета ряда сторонних факторов, главным из которых является контекст. При проверке фактов, проводимой человеком, учитываются исторический и культурный фон, авторитетность упоминающихся в тексте лиц и спикеров и прочие особенности, относящиеся скорее к подаче, чем к непосредственному содержанию текста. Сюда же можно добавить и необходимость верного считывания эмоций, языковых приемов, передающих сарказм и иронию, - областей, в которых человек инстинктивно преуспевает без сознательных усилий. По мнению исследователей, несмотря на значительный прогресс в развитии автоматизированного анализа текстов, ИИ-технологии все еще имеют ограничения в интерпретации и оценке отдельных высказываний [9]. Современные системы искусственного интеллекта успешно идентифицируют и маркируют простые утверждения, однако с трудом справляются с дезинформацией, которая опирается на более тонкие формы выражения

смысла, нежели явное содержание. Кроме того, дополнительную сложность создают языковые барьеры и особенности политических и культурных реалий в отдельно взятой стране. Улучшение необходимых характеристик нейросети в сторону лучшего понимания контекста поможет приблизить автоматизированную проверку информации искусственным интеллектом к человеческой точности. Сегодня прикладываются усилия по разработке систем на основе машинного обучения и алгоритмов обработки естественного языка для обнаружения дезинформации. Однако общим подходом к решению этих проблем на современном этапе является вовлечение человека в процесс анализа текста, особенно при совершенствовании алгоритмов машинного обучения.

Все перечисленные факторы позволяют сделать вывод о будущих перспективах развития нейросетей как инструмента борьбы с информационным фальсификатом и сформулировать основополагающие стратегии их будущего применения.

Первостепенное значение, помимо выявления дезинформации, отводится созданию механизмов минимизации влияния фальшивых новостей. Представляется необходимым разработать алгоритмы прогнозирования появления информационного фальсификата, чтобы получить возможность своевременно предотвратить его распространение. Такой подход может стать наиболее эффективным механизмом борьбы с дезинформацией в долгосрочной перспективе. Помимо этого, закономерным кажется дальнейшее совершенствование нейросетей в процессе контролируемого человеком машинного обучения, качественное улучшение показателей искусственного интеллекта в области идентификации, классификации и маркировки структурированного и неструктурированного контента. Также стоит принимать во внимание тот факт, что на сегодняшний день существует ряд недоступных для нейросетей задач, требующих прямого участия оператора, - это касается точности считывания языковых нюансов в процессе обработки текстов. Нельзя забывать, что и улучшение характеристик

нейросети в процессе контролируемого обучения человеком должно оставаться в высшей степени качественным и беспристрастным.

Подводя итоги, отметим, что помимо автоматизированной проверки приводимых фактов, борьба с дезинформацией в глобальных масштабах требует на современном этапе комплексного подхода, включающего не только использование решений на основе искусственного интеллекта и передовых технологических инструментов, но и непосредственное человеческое участие. Важно продолжать научные исследования, разработку новых и совершенствование существующих алгоритмов, которые помогут автономно определять фальшивые новости. При этом необходимо вести просветительскую работу, объясняя преимущества нейросетей как действенного инструмента против информационного фальсификата, а также разработать соответствующие механизмы контроля. Учитывая ограничения как нейросети, так и человека, можно констатировать, что объединение усилий человеческого и искусственного интеллекта является неоспоримым преимуществом в борьбе с дезинформацией.

Список литературы:

1. Torabi Asr, F, Taboada, M. Big Data and quality data for fake news and misinformation detection // Big Data & Society January–June 2019: 1–14. — URL: https://www.researchgate.net/publication/333326318_Big_Data_and_quality_data_for_fake_news_and_misinformation_detection (дата обращения: 27.03.2024)
2. Кульчицкая Д., Фролова Т. Компьютерные алгоритмы в работе российских информационных агентств (на примере ИА "Интерфакс" и "ТАСС") // Вестник Московского университета. Серия 10. Журналистика. 2020. №1. — URL: https://cyberleninka.ru/article/n/kompyuternye-algoritmy-v-rabote-rossiyskih-informatsionnyh-agentstv-na-primere-ia-interfaks-i-tass (дата обращения: 27.03.2024).

3. Зуйкина К., Соколова Д. Пандемия COVID-19 как медиасобытие: особенности конструирования в социальных медиа // Вестн. Том. гос. ун-та. Филология. 2022. №77. — URL: https://cyberleninka.ru/article/n/pandemiya-covid-19-kak-mediasobytie-osobennosti-konstruirovaniya-v-sotsialnyh-media (дата обращения: 27.03.2024).
4. Объясняем.рф. Официальный интернет-ресурс для информирования о социально-экономической ситуации в России. — URL: https://объясняем.рф/ (дата обращения: 27.03.2024)
5. GPT-4 Technical Report / Open AI. — URL: https://cdn.openai.com/papers/gpt-4.pdf (дата обращения: 27.03.2024).
6. Pause Giant AI Experiments: An Open Letter // Future of Life Institute. — URL: https://futureoflife.org/open-letter/pause-giant-ai-experiments/
7. Резолюция ГА ООН A/78/L.49 от 11 марта 2024 г. — URL: https://undocs.org/Home/Mobile?FinalSymbol=A%2F78%2FL.49&Language=E&DeviceType=Desktop&LangRequested=False (дата обращения: 27.03.2024)
8. Иванова А. Дезинформация в интернете: неизбежная реальность? // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 4, Государство и право: Реферативный журнал. 2023. №3. — URL: https://cyberleninka.ru/article/n/dezinformatsiya-v-internete-neizbezhnaya-realnost (дата обращения: 27.03.2024).
9. Kertysova, K. Artificial Intelligence and Disinformation // Security and Human Rights 29(2018) 55-81. — URL: https://www.researchgate.net/publication/338042476_Artificial_Intelligence_and_Disinformation
10. Международная информационная безопасность: подходы России / А.В.Крутских, Е.А.Зиновьева, В.И.Булва, М.Б.Алборова, Ю.А.Юдина; под ред. А.В.Крутских, Е.С.Зиновьева. — Москва, 2021. — 48 с.

STRATEGIES OF AI-TECHNOLOGIES APPLICATION AGAINST INFORMATION FALSIFICATION

Dmitry A. Chizhov, 2nd year Master's student of the Faculty of International Journalism
MGIMO MFA Russia.
MGIMO 119454, Moscow Vernadsky Prospekt, 76.
E-mail: dmitry4izhov@gmail.com

Abstarct: the issue of using advanced technological tools to combat information falsification is becoming particularly relevant nowadays. This trend is associated with the fact that the latest information technologies are becoming more widespread and occupy an increasingly important place in international communications. In turn, information falsification becomes a serious threat, as it causes a negative impact on public opinion, decision-making processes and the political situation in the country. In this regard, it seems necessary to study the possibilities of using neural networks for timely response to the creation of information falsification and countering its spread in the modern information field.
Key words: neural network, artificial intelligence, GPT-4, fakes, information falsification, disinformation, international communications.