Материал подготовлен ChatGPT
и не является позицией АНО «ИРИ»

Текстовые дипфейки

Что такое текстовые дипфейки?

Текстовые дипфейки (deepfake text или синтетический текст) представляют собой текст, который был сгенерирован алгоритмами искусственного интеллекта так, чтобы выглядеть так, как если бы он был написан человеком.

Эта технология использует методы глубокого обучения (deep learning) для анализа и воспроизведения паттернов человеческого письма, таких как стиль, тон и структура.

Полученный текст можно использовать для создания поддельных новостных статей, постов в социальных сетях, электронных писем и других типов цифрового контента, авторами которого кажутся реальные люди.

Текстовые дипфейки также могут использоваться для выдачи себя за физических лиц и создания поддельных документов, таких как банковские выписки или юридические контракты.

Текстовые дипфейки становятся все более изощренными, из-за чего людям становится трудно отличить подлинный текст от синтетического. Это вызывает обеспокоенность по поводу потенциального воздействия текстовых подделок на общество, включая их использование для распространения дезинформации, совершения мошенничества и манипулирования общественным мнением.

В результате растет интерес к разработке методов обнаружения и предотвращения использования текстовых дипфейков, а также к пропаганде ответственного и этичного использования искусственного интеллекта при генерации текстов.

С какими рисками сопряжено использование текстовых дипфейков?

Текстовые дипфейки могут представлять ряд рисков, в том числе:

1. *Дезинформация.* Текстовые фейки могут использоваться для создания и распространения ложной информации, которая может иметь значительные последствия в реальном мире.

2. *Мошенничество*. Текстовые подделки могут использоваться для того, чтобы выдавать себя за физических лиц и создавать поддельные документы,

такие как банковские выписки, юридические контракты или даже посты в социальных сетях, с целью получения финансовой выгоды.

3. *Кибербезопасность*. Текстовые дипфейки могут использоваться для создания фишинговых электронных писем или других видов кибератак, в которых используются убедительные формулировки, чтобы обманом заставить людей раскрыть конфиденциальную информацию.

4. *Ущерб репутации*. Текстовые подделки могут быть использованы для нанесения ущерба репутации отдельных лиц или организаций путем распространения ложной или порочащей информации.

5. *Политическая манипуляция*. Текстовые дипфейки могут использоваться для манипулирования общественным мнением и влияния на выборы путем создания поддельных новостных статей или постов в социальных сетях, поддерживающих определенную политическую повестку дня.

6. *Юридические последствия*. Текстовые подделки могут привести к юридическим последствиям, таким как диффамация, нарушение авторских прав или даже уголовные обвинения, если они используются в мошеннических целях.

7. *Потеря доверия*. Распространение текстовых подделок может подорвать доверие общественности к онлайн-контенту, что приводит к общему чувству подозрительности и недоверия.

В целом, текстовые дипфейки представляют значительный риск для отдельных лиц, организаций и общества в целом, поскольку способствуют распространению ложной информации, мошенничеству и другим неэтичным практикам.

Важно разработать стратегии по выявлению и предотвращению их использования, а также поощрять ответственное и этичное использование текста, сгенерированного с помощью искусственного интеллекта.

Какие меры позволят снизить риски использования текстовых дипфейков?

Вот некоторые меры по снижению рисков текстовых дипфейков:

1. *Образование и осведомленность:* информирование отдельных лиц и организации о рисках и последствиях текстовых дипфейков, обучение тому, как распознавать их и реагировать на них.

2. *Верификация и проверка фактов:* внедрите системы верификации и проверки фактов, которые могут обнаруживать и предотвращать распространение текстовых дипфейков. Эти системы должны быть способны проверять подлинность источника и содержания информации.

3. *Цифровая аутентификация:* использование технологии цифровой

аутентификации для проверки подлинности онлайн-контента. Эти технологии могут включать цифровые подписи или водяные знаки.

4. *Анализ текста: р*азработка инструментов и технологий, которые могут анализировать язык и структуру текста для выявления потенциальных дипфейков. Эти инструменты должны быть способны обнаруживать аномалии в языке и стиле написания.

5. *Законодательство и нормативные акты:* разработка законодательства , которое будет регулировать создание и распространение поддельных текстов. Оно должно предусматривать штрафы за неправильное использование дипфейков.

6. *Сотрудничество:* содействие сотрудничеству между технологическими компаниями, академическими институтами и правительственными учреждениями в разработке и внедрении решений для обнаружения и предотвращения текстовых дипфейков.

7. *Этические рекомендации:* разработка этических рекомендаций по созданию и использованию дипфейков. Эти руководящие принципы должны касаться таких вопросов, как согласие, точность, указание авторства и конфиденциальность.

8. *Медиаграмотность:* пропаганда медиаграмотности и навыковкритического мышления среди людей, чтобы помочь им идентифицировать сгенерированный текст и реагировать на него.

В целом, снижение рисков текстовых дипфейков требует многостороннего подхода, который включает образование, технологии, регулирование и сотрудничество. Реализуя эти меры, мы можем уменьшить влияние текстовых дипфейков на отдельных лиц, организации и общество в целом.

Примеры текстовых дипфейков.

Вот несколько более конкретных примеров текстовых дипфейков.

Новостные статьи, генерируемые GPT-3: GPT-3 – это мощная языковая модель, которая может генерировать текст, похожий на человеческий. Исследователи продемонстрировали, как GPT-3 может быть использован для создания убедительных новостных статей, содержащих ложную информацию. Например, исследовательская группа под названием OpenAI создала поддельную новостную статью о ядерной катастрофе, которая смогла обмануть некоторых читателей, заставив их думать, что это реально.

Фишинговые электронные письма, генерируемые искусственным интеллектом: инструменты на базе искусственного интеллекта можно

использовать для создания фишинговых электронных писем, имитирующих стиль письма надежного отправителя, такого как банк или платформа социальных сетей. Эти электронные письма могут быть адаптированы к конкретным получателям, что делает их более убедительными. Например, фишинговое электронное письмо, сгенерированное искусственным интеллектом, может быть адресовано получателю по имени и содержать сведения об его учетной записи, чтобы оно выглядело более легитимным.

Поддельные отзывы о ресторанах: Исследователи показали, как сгенерированные искусственным интеллектом отзывы можно использовать для манипулирования онлайн-рейтингами ресторанов или других предприятий. Например, инструмент под названием GROVER может генерировать убедительные отзывы о ресторанах, которые трудно отличить от реальных отзывов, написанных людьми.

Дипфейки общественных деятелей: В то время как большинство людей ассоциируют дипфейки с манипулируемыми изображениями или видео, также возможно создавать текстовые дипфейки, которые выдают себя за общественных деятелей. Например, исследователи использовали GPT-2 для генерации твитов, имитирующих стиль письма Дональда Трампа, которые могут быть использованы для распространения ложной информации или создания путаницы в социальных сетях.

Научные статьи, созданные с помощью искусственного интеллекта: Исследователи продемонстрировали, как искусственный интеллект может быть использован для создания убедительных научных работ, которые являются плагиатом из существующих работ. Например, группа исследователей создала инструмент под названием SCIgen, который мог генерировать статьи по информатике, которые были приняты на конференциях и в журналах, несмотря на то, что содержали бессмысленный текст.

Примеры реализации рисков использования текстовых дипфейков.

Вот несколько реальных примеров рисков, связанных с использованием текстовых дипфейков.

Дезинформация: Текстовые дипфейки могут быть использованы для создания убедительных новостных статей, постов в социальных сетях и другого письменного контента, распространяющего ложную информацию или дезинформацию по существу. В 2020 году группа исследователей создала текстогенерирующую модель искусственного интеллекта, которая смогла писать убедительные фейковые новостные статьи о пандемии COVID-19.

Исследователи использовали эту модель, чтобы подчеркнуть потенциальные риски использования текста, сгенерированного искусственным интеллектом, для распространения ложной информации. (Источник: https://www.technologyreview.com/2020/06/04/1002474/ai-fake-text-gpt-2-dangerous /)

В 2020 году на PRNewswire был опубликован ложный пресс-релиз, в котором утверждалось, что Walmart начнет принимать Litecoin в качестве формы оплаты. Пресс-релиз был текстовой подделкой, и хотя он был быстро опровергнут, это вызвало краткий скачок цены лайткоина. (Источник: https://www.bbc.com/news/technology-58323858)

Фишинговые мошенничества: Текстовые подделки могут использоваться в фишинговых атаках или для распространения вредоносного ПО. В 2020 году текстовая подделка была использована в фишинг-атаке на британскую энергетическую фирму. Злоумышленник использовал текстовую подделку, чтобы выдать себя за генерального директора компании и убедить сотрудника перевести деньги на мошеннический счет. (Источник: https://www.bbc.com/news/technology-53432590)

Киберзапугивание: Текстовые подделки могут использоваться для создания поддельных постов в социальных сетях или сообщений, которые могут быть использованы для киберзапугивания или домогательств. В 2019 году группа исследователей создала текстогенерирующую модель искусственного интеллекта, которая смогла генерировать поддельные обзоры Yelp, которые было трудно отличить от реальных отзывов. Исследователи отметили, что этот тип технологии может быть использован для киберзапугивания или для нанесения ущерба репутации бизнеса. (Источник: https://www.businessinsider.com/ai-is-being-used-to-fake-reviews-on-yelp-and-amazon-2019-10)

Эти примеры иллюстрируют некоторые риски, связанные с использованием текстовых подделок, включая возможность распространения дезинформации, совершения мошенничества и участия в киберзапугивании или домогательствах.

Комментарий АНО «ИРИ»:

Представленная GPT информация о текстовых дипфейках в целом отвечает действительности и информирует пользователей разной степени осведомленности о существующей проблематике.

Риски, связанные с использованием дипфейков, определены корректно и отвечают действительности.

Интересны примеры из реальной жизни, которые сервис определил как текстовые подделки.

Поскольку ChatGPT генерирует текст без ссылок на используемые источники, у сервиса было запрошено, откуда была взята информация о реальных случаях текстовых дипфейков. Некоторые из ссылок оказались недействительными и найти первоисточник не удалось.

Так, из всех представленных сервисом ссылок работают только ссылки на статьи Гровера и Вашингтонского университета. Остальные ссылки некорректны, хотя приведенные сервисом примеры действительно имели место быть и проверяются другими достоверными источниками.

Однако имеются некоторые ошибки в ряде фактов, а именно использование текстового дипфейка в фишинговой атаке на британскую энергетическую фирму в 2020 году. Во-первых, такая атака произошла в 2019 году[1]*. Во-вторых, использовался не текстовый дипфейк, а подделанный голос.*

Ложный пресс-релиз об принятии Litcoin сетью Walmart был опубликован не в 2020 году, а в 2021.

[1] https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=4195fae12241