ПРИМЕНЕНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В РАМКАХ БОРЬБЫ С ДЕЗИНФОРМАЦИЕЙ: СОВРЕМЕННЫЕ МЕТОДЫ И РЕШЕНИЯ

Орлов Д. В., Коростелёв А. Д., Марков А. Н.
Центр информатизации и инновационных разработок, кафедра информатики,
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: {d.orlov, a.korostelev, a.n.markov}@bsuir.by

Данный доклад посвящен актуальной проблеме распространения дезинформации в цифровом пространстве и анализу роли искусственного интеллекта как инструмента её создания и противодействия. Рассматриваются современные технологии генерации синтетического контента с использованием нейронных сетей, подходы к его выявлению и автоматизированные механизмы противодействия. Отдельное внимание уделено этическим и правовым аспектам использования подобных технологий.

Введение

Цифровые технологии играют важную роль в формировании современного общества, делая социальные сети, новостные платформы и мессенджеры ключевыми каналами распространения информации. Вместе с ростом доступности информации обострилась проблема её достоверности: фейковые новости, манипулятивные тексты, синтетические мультимедийные материалы становятся всё более массовыми [1].

Искусственный интеллект играет в этой сфере двойственную роль. С одной стороны, генеративные модели позволяют создавать контент, который сложно отличить от настоящего, что облегчает распространение ложной информации и манипуляцию общественным мнением. С другой стороны, современные методы анализа текста, изображений, аудио и сетевых взаимодействий предоставляют возможности для автоматического выявления и фильтрации ложного контента.

Таким образом, перед исследователями и разработчиками стоит комплексная задача: с одной стороны, изучить и понять механизмы создания дезинформации с помощью ИИ, с другой – разработать эффективные методы её обнаружения и контрмеры, учитывая технологические, этические и правовые аспекты.

I. Генерация дезинформации с помощью ИИ

Современные генеративные модели позволяют создавать фейковые тексты, изображения, видео и аудио с высокой степенью правдоподобия. Большие языковые модели (LLM) генерируют тексты, имитирующие стиль журналистских статей, научных публикаций или социальных постов, и могут содержать манипулятивные аргументы, что делает их эффективными в распространении нодостоверной информации. Генеративные нейросети (GAN, diffusion-модели) создают реалистичные изображения и видеоролики, включая дипфейки, где лица людей и их мимика выглядят естественно и практически неотличимы от настоящих [1].

Технологии синтеза голоса позволяют имитировать интонацию, тембр и стиль речи конкретного человека, что используется для подделки интервью или голосовых сообщений. Доступность облачных сервисов и open-source библиотек снижает порог входа для злоумышленников: создание и распространение синтетического контента стало возможным даже для непрофессионалов. При этом генеративные модели могут комбинировать текст, изображение и аудио, создавая мультимедийные фейки, которые сложно обнаружить традиционными методами. Масштабирование таких систем позволяет быстро распространять дезинформацию, что усиливает её влияние на общественное мнение и затрудняет своевременное выявление.

II. Методы обнаружения дезинформации

Одним из ключевых направлений является использование NLP-подходов (Natural Language PRocessing) для анализа текстового контента. Модели обработки естественного языка классифицируют тексты на категории «правда» или «ложь», выявляют стилистические аномалии и сопоставляют утверждения с источниками. Современные трансформеры, такие как BERT (Bidirectional Encoder Representations from Transformers) и RoBERTa (Robustly Optimized BERT Approach), позволяют извлекать ключевые утверждения и проводить их верификацию с помощью базы фактов. Основное ограничение NLP-подходов заключается в сложности интерпретации сарказма, цитат или контекста, что может приводить к ошибочным классификациям.

Другой эффективный метод – использование графовых моделей и анализ сетевой структуры распространения информации. Анализ сетей помогает обнаруживать закономерности координированных действий, активность ботов и подозрительные группы распространения информации [2]. Графовые нейронные сети помогают

обнаружать аномалии в связях между пользователями и публикациями, всплески репостов и аномальные маршруты контента. Основные проблемы применения графовых моделей включают необходимость доступа к сетевым данным платформ и необходимость учитывать естественные вирусные эффекты, чтобы не путать их с координированной дезинформацией.

Направление визуального анализа фокусируется на выявлении дипфейков и синтетических изображений. Сверточные и временные нейросети ищут артефакты сжатия, несоответствия мимики, освещения и аномалии в частотной области. Мультимодальные проверки, которые сравнивают видео и аудио, повышают точность выявления синтетических материалов. Ограничение этого подхода в том, что генеративные модели становятся всё более реалистичными, что снижает эффективность традиционных признаков детекции [3].

На практике наиболее эффективными являются комплексные системы, которые объединяют текстовый, визуальный и сетевой анализ с проверкой метаданных. Наборы моделей и методы cross-checking позволяют снизить количество ложных срабатываний, повышая точность детекции. Основной недостаток таких систем – высокая сложность реализации и значительные вычислительные ресурсы, необходимые для обработки больших объёмов данных.

III. Контрмеры

Одним из методов противодействия синтетическому контенту является внедрение водяных знаков (watermarking). С помощью невидимых или видимых меток можно отслеживать источник сгенерированного ИИ-контента и идентифицировать его происхождение. Применяются криптографические методы и стеганографические подходы, которые делают метки устойчивыми к базовым изменениям файла. Основное ограничение водяных знаков заключается в том, что они могут быть удалены при конвертации, сжатии или повторном распространении контента [4].

Хранение информации о происхождении контента и его цепочке изменений (provenance) является ещё одним важным инструментом. Метаданные позволяют установить подлинность файла, отслеживать источники и выявлять подделки. Технологии цифровой подписи и хэширования уменьшают возможность фальсификации, однако часто метаданные теряются при скачивании, пересылке или преобразовании контента, что ограничивает их практическую эффективность.

Полуавтоматические системы с участием человека (human-in-the-loop) сочетают автоматическую фильтрацию с экспертной проверкой. Алгоритмы предварительно анализируют контент и выделяют подозрительные материалы, после чего эксперт принимает окончательное решение. Такой подход снижает риск ложных блокировок, улучшает обучение моделей за счёт обратной связи и повышает точность дикции. Главный недостаток – плохая масштабируемость при больших объёмах данных и высокая потребность в человеческих ресурсах [2].

Регуляторы и крупные платформы внедряют правила маркировки синтетического контента, а также программы прозрачности и fact-checking, чтобы систематизировать борьбу с дезинформацией. Такие инициативы помогают выстраивать процессы обнаружения и реагирования на фейки, однако стандарты и подходы сильно различаются между странами и компаниями. Это создает фрагментированную картину регулирования и требует согласования технических и правовых норм для повышения эффективности [4].

IV. Этические вопросы

Автоматические системы детекции несут риск ложных срабатываний и цензуры. Избыточная автоматизация может подавлять свободу слова и использоваться для контроля информации. Необходимы прозрачность алгоритмов, процедуры апелляции и участие независимых аудиторов для минимизации побочных эффектов.

V. Выводы

ИИ не только способствует созданию дезинформации, но и предоставляет инструменты для её обнаружения и борьбы с ней. Перспективным является развитие гибридных систем, совмещающих машинное обучение, watermarking и участие человека. Эффективная борьба с дезинформацией требует внедрения технологических решений, правового регулирования и повышения цифровой грамотности пользователей.

VI. Список литературы

1. Баяк, Е. И. Применение Big Data для анализа социальных сетей / Е. И. Баяк, С. Н. Нестеренков, Д. А. Жалейко // BIG DATA и анализ высокого уровня = BIG DATA and Advanced Analytics : сб. науч. ст. IX Междунар. науч.-практ. конф. В 2 ч. Ч. 1 (Республика Беларусь, Минск, 17–18 мая 2023 года) / редкол. : В. А. Богуш [и др.]. – Минск : БГУИР, 2023. – С. 348–350.
2. Половой, А. А. Когнитивные аспекты взаимодействия человека с нейросетевыми системами / А. А. Половой, С. Н. Нестеренков // Информатика : сборник трудов международной молодежной школы «Инженерия – XXI» (Российская Федерация, г. Новороссийск, 15-18 апреля 2025 г.) / редкол : И. В. Чистяков [и др.]. – Новороссийск : БГТУ им. В. Г. Шухова, 2025. – С. 224-226.
3. Mirsky, Y., Lee, W. The Creation and Detection of Deepfakes: A Survey // ACM Computing Surveys. – 2021. – Vol. 54, № 1. – P. 1-41
4. European Commission. Code of Practice on Disinformation [Electronic resource] / European Commission. – Brussels, 2023. – Mode of access: `https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation` . – Date of access: 26.09.2025.