DOI 10.37882/2223-2966.2025.01.16

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В КИБЕРУГРОЗАХ: ВОЗМОЖНОСТИ И ВЫЗОВЫ ИСПОЛЬЗОВАНИЯ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ФИШИНГОВЫХ АТАК И ИХ ДЕТЕКТИРОВАНИЯ

ARTIFICIAL INTELLIGENCE IN CYBER THREATS: OPPORTUNITIES AND CHALLENGES OF USING NEURAL NETWORKS FOR PHISHING ATTACKS AND THEIR DETECTION

N. Donskikh

Summary. This article examines the impact of artificial intelligence, in particular large language models, on the evolution of phishing attacks, as well as the possibilities of countering these threats using classification neural networks. It considers the methods that attackers use to create personalized and convincing phishing messages and analyzes the effectiveness of neural networks in their detection and prevention. Particular attention is paid to the importance of balancing the development of AI technologies and their regulation to ensure cybersecurity. In conclusion, technical, legal, and ethical measures aimed at regulating and safely using these technologies are proposed.

Keywords: artificial intelligence, large language models, phishing attacks, neural networks, classification models, cybersecurity, AI regulation, social engineering, personalization of attacks, AI technologies.

Донских Никита Игоревич
Аспирант, Финансовый университет при Правительстве РФ, г. Москва
Nikdonskikh@gmail.com

Аннотация. В статье исследуется влияние искусственного интеллекта, в частности больших языковых моделей, на эволюцию фишинговых атак, а также возможности противодействия этим угрозам с помощью классификационных нейронных сетей. Рассматриваются методы, которые злоумышленники используют для создания персонализированных и убедительных фишинговых сообщений, а также анализируется эффективность нейросетей в их выявлении и предотвращении. Особое внимание уделяется важности баланса между развитием ИИ-технологий и их регулированием для обеспечения кибербезопасности. В заключении предложены технические, правовые и этические меры, направленные на регулирование и безопасное использование этих технологий.

Ключевые слова: искусственный интеллект, большие языковые модели, фишинговые атаки, нейронные сети, классификационные модели, кибербезопасность, регулирование ИИ, социальная инженерия, персонализация атак, технологии ИИ.

Введение

Современные технологии искусственного интеллекта (ИИ) открывают перед человечеством широкие перспективы, позволяя автоматизировать процессы, улучшать качество услуг и создавать инновационные продукты. Однако, вместе с положительными изменениями, нейронные сети, особенно крупные языковые модели, становятся инструментом, который могут использовать злоумышленники. Одной из наиболее значимых угроз сегодня является применение ИИ для реализации фишинговых атак.

Фишинг, основанный на использовании социальных приемов манипуляции, ранее часто сводился к рассылке стандартных сообщений, которые легко идентифицировались благодаря грамматическим ошибкам и небрежному оформлению [1]. Но с развитием языковых моделей, таких как GPT, стало возможным создавать тексты, которые выглядят настолько правдоподобно, что их трудно отличить от работы человека. Это привело к значительному увеличению точности и эффективности фишинговых атак, усложняя их выявление.

Параллельно с этим, прогресс в создании классификационных моделей на основе нейронных сетей открывает новые возможности для эффективного выявления фишинговых атак. Такие модели способны анализировать текстовые сообщения, оценивая их подлинность, что позволяет минимизировать риск утечек данных. Однако быстрая адаптация злоумышленников к новым методам требует от разработчиков систем безопасности постоянной работы над усовершенствованием своих решений [2].

В статье рассматриваются перспективы использования крупных языковых моделей для реализации фишинговых атак, а также способы их обнаружения с помощью классификационных нейронных сетей. Особое внимание уделяется необходимости нахождения баланса между развитием технологий искусственного интеллекта и обеспечением их безопасного применения.

Результаты исследования

Фишинг, как форма социальной инженерии, нацелен на обман пользователей с целью получения их конфи-

денциальных данных, таких как пароли, информация о банковских картах или доступ к системам. Ранее злоумышленники часто использовали стандартные шаблоны сообщений, которые легко идентифицировались благодаря орфографическим ошибкам и явным признакам подделки. Однако появление крупных языковых моделей, например GPT, существенно изменило подход к таким атакам [3].

Языковые модели, обученные на огромных объемах текстовых данных, способны создавать связные, грамматически корректные и стилистически выверенные тексты. Анализ данных из открытых источников, таких как социальные сети, позволяет злоумышленникам генерировать сообщения, адресованные конкретным людям. Такие письма могут содержать персональную информацию, что значительно повышает доверие жертвы. Модели способны воспроизводить тон и стиль официальной корреспонденции, например писем от банков, IT-компаний или других организаций, включая использование профессиональной терминологии и корпоративной лексики. Благодаря ИИ можно быстро генерировать сотни или даже тысячи уникальных сообщений, что затрудняет их выявление спам-фильтрами и увеличивает шансы на успех атаки [4].

Злоумышленник может воспользоваться языковой моделью для создания убедительного письма от имени службы поддержки популярного интернет-магазина. В таком письме будет содержаться правдоподобное объяснение, почему пользователю нужно срочно перейти по ссылке, например, чтобы подтвердить покупку или восстановить доступ к своему аккаунту [5].

Особенности использования GPT в фишинговых атаках:

1. Автоматизация переводов — модели ИИ дают возможность организовывать фишинговые кампании на нескольких языках, адаптируя их под конкретные регионы и аудитории.
2. Контекстная генерация контента — вводя ключевые данные, такие как имя адресата или название компании, злоумышленники могут получать тексты, которые максимально соответствуют заданной ситуации.
3. Обход фильтров — благодаря умению генерировать тексты без характерных ключевых слов, языковые модели помогают обходить фильтры, настроенные на выявление спама.

Исследования показывают, что языковые модели активно применяются для создания более изощрённых атак. Например, письма с ложными уведомлениями о взломе аккаунта или поддельными инструкциями по обновлению безопасности становятся всё более распространёнными. Актуальные темы, такие как COVID-19, часто используются для разработки фишинговых кампаний, в которых ИИ создаёт тексты, нацеленные на эксплуатацию страхов и интересов жертв [6].

Применение языковых моделей делает фишинг особенно опасным, поскольку создаваемые тексты трудно отличить от сообщений, составленных человеком. Это поднимает необходимость разработки новых методов защиты, которые будут рассмотрены в следующем разделе.

С увеличением сложности фишинговых атак, вызванным применением больших языковых моделей, традиционные средства защиты, такие как фильтрация ключевых слов или проверка отправителей, теряют свою эффективность. Чтобы противостоять этим новым вызовам, разработчики всё чаще используют классификационные нейронные сети, которые способны анализировать множество параметров сообщений и оценивать их подлинность [7].

Классификационные нейронные сети основаны на методах машинного обучения, что позволяет им адаптироваться к постоянно изменяющимся угрозам. Среди ключевых преимуществ таких технологий можно выделить:

Глубокий текстовый анализ — эти модели учитывают синтаксические, семантические и стилистические аспекты текста, позволяя выявлять даже скрытые признаки фишинга.

Анализ метаданных — помимо текста, нейронные сети рассматривают дополнительные параметры, такие как IP-адрес отправителя, время отправки, структура заголовков и вложений, что помогает формировать более точную оценку.

Обучение на больших массивах данных — такие модели проходят обучение на миллионах примеров, включая как безопасные, так и фишинговые сообщения, что позволяет эффективно различать их.

Выявление аномалий — нейронные сети способны обнаруживать подозрительное поведение, даже если оно не соответствует известным паттернам атак, что особенно важно в условиях постоянной эволюции угроз [8].

Алгоритм работы классификационной модели представлен на рисунке 1.

Классификационные модели находят широкое применение в различных сферах кибербезопасности. Например, почтовые сервисы, такие как Gmail и Outlook, используют нейронные сети для анализа входящих сообщений, что позволяет автоматически перемещать

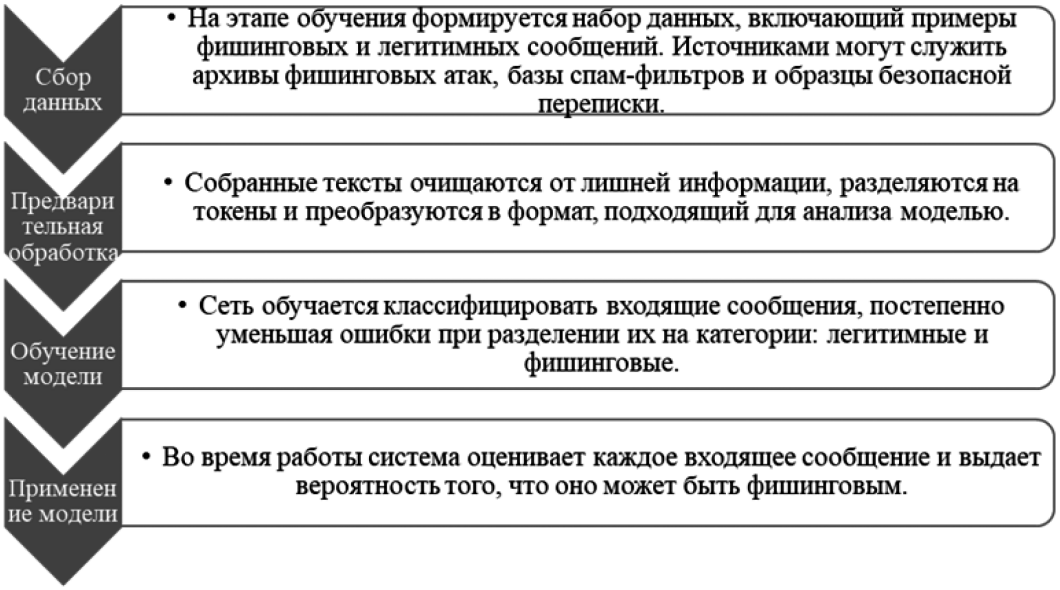

Рис. 1. Алгоритм работы классификационной модели

подозрительные письма в папку «Спам» и защищать пользователей от потенциальных угроз. Кроме того, такие модели применяются для анализа URL-адресов, проверяя, ведут ли ссылки на поддельные сайты или представляют другую опасность [10]. В случае выявления фишинговой активности системы могут оперативно уведомлять пользователей или администраторов, предотвращая дальнейшее распространение угроз.

Несмотря на высокую эффективность, классификационные модели имеют ряд существенных ограничений. Одной из главных проблем являются фальшивые срабатывания, когда легитимные сообщения ошибочно классифицируются как фишинговые, что может создавать неудобства и нарушать рабочие процессы. Кроме того, злоумышленники активно адаптируют свои методы, используя технологии ИИ для обхода защитных систем, что вынуждает разработчиков регулярно обновлять модели [9]. Также значительным препятствием остаются высокие вычислительные затраты, так как анализ и обработка больших объемов данных требуют значительных ресурсов.

Будущее классификационных моделей связано с интеграцией ИИ и сопутствующих технологий, таких как гибридные системы, которые объединяют нейронные сети с эвристическими методами и базами данных угроз для повышения эффективности защиты. Перспективным направлением является внедрение самообучающихся алгоритмов, способных автоматически обновляться на основе новых данных о фишинговых атаках. Также особое внимание уделяется учёту контекста, что позволит моделям анализировать сообщения в рамках переписки пользователя и точнее выявлять угрозы. Уже сегодня классификационные нейронные сети играют важнейшую роль в противодействии фишингу, а их постоянное совершенствование станет залогом повышения кибербезопасности в условиях растущей сложности атак [11].

Вывод

Бурное развитие технологий искусственного интеллекта, особенно больших языковых моделей, значительно изменило ландшафт киберугроз, делая фишинговые атаки более убедительными и масштабируемыми. Способность таких моделей создавать персонализированные и стилистически выверенные тексты бросает вызов традиционным методам защиты.

В то же время классификационные нейронные сети показывают высокую эффективность в борьбе с этими угрозами, анализируя текстовые и метаданные сообщений, выявляя признаки фишинга и приспосабливаясь к новым тактикам злоумышленников. Однако искусственный интеллект остается двусторонним инструментом, используемым как для защиты, так и для совершения атак.

Для обеспечения надежной кибербезопасности требуется системный подход, включающий технические решения, правовые нормы и этические стандарты. Разработка механизмов регулирования ИИ должна минимизировать риски его злоупотребления, одновременно поддерживая инновационное развитие.

Современные вызовы киберугроз требуют объединения усилий разработчиков, исследователей, государственных институтов и пользователей. Только совместные действия, основанные на диалоге и принципах ответственного использования технологий, позволят найти баланс между прогрессом и защитой данных в эпоху глобальной цифровизации.

ЛИТЕРАТУРА

1. Белокопытов А.С. Применение искусственного интеллекта в сферах кибербезпасности / А.С. Белокопытов, С.С. Яковлева // Фундаментальные и прикладные научные исследования в современном мире // Сборник научных статей по материалам II Международной научно-практической конференции, Уфа, 09 июня 2023 года. Том Часть 3. Уфа: Общество с ограниченной ответственностью «Научно-издательский центр «Вестник науки». 2023. С. 263–271.
2. Брынза И.Г. Искусственный интеллект и кибербезопасность: вызовы и перспективы / И.Г. Брынза // Кибербезопасность и информационные технологии. 2019. №2(15). С. 45–49.
3. Иванов А.В. Искусственный интеллект в задачах кибербезопасности. / А.В. Иванов // Информационно-управляющие системы. 2021. №1. С. 69–73.
4. Кузнецов Н.Н. Применение методов искусственного интеллекта для повышения эффективности систем информационной безопасности. / Н.Н. Кузнецов // Защита информации. 2017. № 4. С. 4–13.
5. Муковнин Г.М. Анализ методов защиты от утечек данных в корпоративных сетях / Г.М. Муковнин // Цифровые системы и модели: теория и практика проектирования, разработки и применения: Материалы национальной (с международным участием) научно-практической конференции, Казань, 10–11 апреля 2024 года. Казань: Казанский государственный энергетический университет. 2024. С. 1347–1350.
6. Мухамадиева К.Б. Обзор методов обнаружения фишинговых атак на основе искусственного интеллекта / К.Б. Мухамадиева, Б.Б. Муминов // Вестник Донецкого национального университета. Серия Г: Технические науки. 2021. № 4. С. 37–45.
7. Николаев А.А. Применение искусственного интеллекта в системах кибербезопасности. / А.А. Николаев, Е.А. Степанов // Проблемы информационной безопасности. 2018. Т. 1. № 1. С. 15–21.
8. Петров Д.А. Искусственный интеллект в задачах прогнозирования уязвимостей информационной безопасности / Д.А. Петров // Системное администрирование и информационная безопасность. 2020. Т. 17. № 2. С. 52–58.
9. Ручай А.Н. Методы машинного обучения и искусственного интеллекта в сфере информационной безопасности: анализ современного состояния и перспективы развития / А.Н. Ручай, И.В. Токарев, А.С. Грибачев // Вестник УрФО. Безопасность в информационной сфере. 2022. № 4(46). С. 76–87.
10. Соколов А.С. Применение искусственного интеллекта в задачах обнаружения и предотвращения кибератак / А.С. Соколов // Труды Института системного анализа РАН. 2019. Т. 76. № 1. С. 148–159.
11. Шарипов Р.Р. Применение машинного обучения и искусственного интеллекта в кибербезопасности. / Р.Р. Шарипов, М.Г. Амиров // Компьютерные инструменты в образовании. 2020. Т. 13. № 2. С. 72–80.

Журнал «Современная наука: актуальные проблемы теории и практики»