DOI 10.37882/2223-2966.2025.06.20

АНАЛИЗ ПАТТЕРНОВ НЕЙРОННОЙ СЕТИ НА ОСНОВЕ ЯЗЫКОВОЙ МОДЕЛИ GPT-3.5 ВО ВРЕМЯ ПРИМЕНЕНИЯ ЗЛОУМЫШЛЕННИКАМИ В ЦЕЛЯХ ФИШИНГА

ANALYSIS OF NEURAL NETWORK PATTERNS IN THE GPT-3.5 LANGUAGE MODEL DURING ITS USE BY MALICIOUS ACTORS FOR PHISHING ATTACKS

V. Dmitrieva
I. Shatov
D. Batyanova
V. Fedorov

Summary. This article is dedicated to the investigation of the capabilities of a neural network based on the OpenAI GPT-3.5 language model in the context of its potential use by malicious actors for generating phishing emails. Particular attention is given to identifying patterns and specific characteristics that emerge when this model is employed in phishing attacks. The aim of the study is to analyze the structure and features of phishing messages created using the neural network, as well as to determine approaches for imitating typical phishing templates in both formal (business) and informal (friendly) styles.

The research involved 43 experiments, each consisting of a series of messages generated by the neural network. The analysis focused on criteria such as the reasons for writing the email, subject lines, sender and recipient identities, and communication style. The research methodology included a detailed examination of the text, formatting practices, and the model's ability to adapt its writing style based on predefined parameters.

The results demonstrated that the neural network is capable of effectively generating emails in a formal business style; however, in the informal style, messages often lose naturalness, which reduces their credibility. In conclusion, the article highlights the importance of continued research into such technologies — not only in terms of their advancement but also in the broader context of information security.

Keywords: phishing, neural networks, OpenAI GPT-3.5, natural language processing, social engineering.

Дмитриева Виктория Александровна
Российский технологический университет, МИРЭА
a79258487537@gmail.com
Шатов Игорь Алексеевич
Российский технологический университет МИРЭА
shat.igor2017@gmail.com
Батьяновна Дарья Денисовна
Российский технологический университет МИРЭА
vin508005@gmail.com
Федоров Вадим Валерьевич
Старший преподаватель, Российский технологический университет МИРЭА
fedorov_v@mirea.ru

Аннотация. Статья посвящена исследованию возможностей нейронной сети на базе языковой модели OpenAI GPT-3.5 при её потенциальном использовании злоумышленниками для создания фишинговых писем. Особое внимание уделяется выявлению паттернов и особенностей, возникающих при использовании данной модели для фишинговых атак. Целью исследования является анализ структуры и характеристик фишинговых сообщений, созданных с помощью нейронной сети, а также определение подходов к имитации типовых шаблонов фишинга в деловом и дружеском стилях. В ходе исследования было проведено 43 эксперимента, каждый из которых включал серию сообщений, сгенерированных нейросетью. Для анализа использовались такие критерии, как причины написания письма, темы, отправители и получатели сообщений, а также стиль общения. Методы исследования включают подробный анализ текста, его форматирования и способности нейросети адаптировать стиль письма в зависимости от заданных параметров. Результаты показали, что нейронная сеть эффективно генерирует письма в деловом стиле, однако в дружеском стиле сообщения часто теряют естественность, что снижает их правдоподобность. В заключение статья подчеркивает важность дальнейшего изучения подобных технологий не только с точки зрения их развития, но и в контексте информационной безопасности.

Ключевые слова: фишинг, нейронные сети, Open AI Gpt 3.5, обработка естественного языка, социальная инженерия.

Введение

Генеративно-преобразующие сети (ГПС) представляют собой мощный и высокоэффективный класс нейронных сетей, предназначенных для создания новых данных, таких как текст, изображения, музыка и другие виды контента. Эти сети обучаются на огромных объемах информации и могут генерировать контент, который соответствует определённым закономерностям или стилю, представленному в исходных данных. В последние годы ГПС приобрели широкую популярность, в том числе благодаря их выдающимся возможностям в области обработки естественного языка (ОЕЯ), что открывает новые горизонты для применения искусственного интеллекта в различных сферах. Согласно данным анализа востребованности ИИ-инструментов, одним из самых узнаваемых и применяемых решений является ChatGPT от компании OpenAI. Его отметили 93 % респондентов

как инструмент, способный генерировать текст в диалоговом режиме. Это подтверждает широкое распространение и потенциальную уязвимость подобных моделей в условиях киберугроз [1, с. 3] [8].

Одним из ярких примеров популярных моделей является OpenAI GPT-3.5, которая продемонстрировала исключительные способности к обработке текста. Модель способна эффективно выполнять задачи, связанные с генерацией текста, переводом между языками, а также обеспечивать точные и содержательные ответы на вопросы. GPT-3.5 использует сложные архитектуры трансформеров и обучена на больших объемах текстовых данных, что позволяет ей учитывать контекст и структуру языка, обеспечивая высокую степень адекватности и точности в выполнении поставленных задач. Способность GPT-3.5 к генерации правдоподобных текстов, а также к автоматизации различных речевых задач, привела к появлению новых угроз, в частности в сфере кибербезопасности. Исследователи отмечают, что искусственный интеллект, помимо очевидной пользы, может быть использован и для совершения преступлений, включая фишинг, мошенничество и социальную инженерию, особенно при наличии у злоумышленников доступа к открытым ИИ-сервисам [2, с. 342] [9] [10].

Материалы и методы

В данном исследовании было проведено 43 эксперимента, каждый из которых включал серию сообщений (не более трех) с целью создания нейронной сетью правдоподобных фишинговых писем, соответствующих заранее заданным критериям. Важно отметить, что каждый эксперимент начинался с нового блока переписки, что исключало наличие контекста, созданного предыдущими сообщениями.

Критерии, задаваемые чату: причины написания фишингового письма, тема письма, отправитель и получатель, стиль.

Причины написания фишингового письма: среди заданных причин были указаны цели фишинга, написание статьи и дипломной работы.

Темы писем: задачи, связанные с созданием фишинговых ссылок для различных категорий пользователей: фишинговая ссылка для крупной компании, для учебного заведения (вуза) и для друга.

Отправители: получателями фишинговых писем были определены директор компании, штатный сотрудник, преподаватель, а также друг.

Получатель: директор компании, штатный сотрудник, преподаватель, друг.

Стили: исследуемые стили включали деловой и дружеский, которые представляют собой два наиболее часто используемых стиля для фишинговых сообщений.

Методология исследования заключалась в подробном анализе нейронной сети с акцентом на ее способность адаптировать текст в зависимости от заданных параметров. Каждый блок сообщений был тщательно сконструирован для того, чтобы исследовать, насколько нейронная сеть способна имитировать типичные паттерны фишинговых писем, сохраняя заданные критерии. Важным аспектом было соблюдение формата письма, а также наличие встроенных механизмов форматирования. Авторы работ по обнаружению аномалий в электронной переписке подчёркивают, что фишинговые письма могут содержать как низкоуровневые сенсорные аномалии, так и высокоуровневые семантические ошибки. Анализ показывает, что даже с учётом корректного форматирования, подобные сообщения могут отклоняться от норм естественной речи, что делает их потенциально обнаруживаемыми алгоритмами машинного обучения [3, с. 1353].

Литературный обзор

Генеративно-преобразующие сети (ГПС) и их использование в области кибербезопасности стали объектом многочисленных исследований. Исследования показывают, что такие технологии могут значительно повысить эффективность как в генерации контента, так и в создании угроз. Согласно ряду авторов, искусственный интеллект, включая ГПС, может быть использован для автоматизации многих процессов, включая создание фишинговых атак, что делает его опасным инструментом для злоумышленников. Одним из ярких примеров такого использования является языковая модель GPT-3.5, разработанная компанией OpenAI, которая, несмотря на свои выдающиеся характеристики, также представляет угрозу в контексте кибербезопасности. Эффективность GPT-3.5 в генерации текста делает её идеальным инструментом для фишинговых атак, где важно создавать правдоподобные сообщения для обмана пользователей [2, с. 342] [9].

Известно, что фишинг — это метод киберпреступности, при котором злоумышленники пытаются обманом получить конфиденциальную информацию пользователя, включая пароли, номера кредитных карт и другие личные данные. В последние годы фишинговые атаки стали значительно более изощрёнными благодаря использованию ИИ и нейронных сетей. В частности, использование ИИ для генерации фишинговых писем приводит к созданию сообщений, которые трудно отличить от легитимной переписки. Это подтверждается исследованиями, которые отмечают, что ИИ способен не только генерировать правдоподобные тексты, но и адапти-

ровать их под конкретные задачи и целевые группы [4, с. 133] [6, с. 3].

Современные работы по обнаружению фишинга акцентируют внимание на использовании методов машинного обучения для автоматического распознавания фишинговых писем [5, с. 164]. Несмотря на высокое качество текстов, генерируемых ИИ, исследования показывают, что такие письма часто можно распознать благодаря определённым аномалиям, как в контексте, так и в структуре сообщения [7, с. 236]. Однако, несмотря на успехи в этой области, борьба с фишингом остаётся актуальной проблемой, поскольку с развитием технологий фишинговые атаки становятся всё более сложными и трудноопределимыми.

Результаты

В ходе эксперимента было создано 43 фишинговых письма, которые были сгенерированы с использованием нейронной сети GPT-3.5. Каждый эксперимент включал серию сообщений (не более трёх), при этом каждый новый эксперимент начинался с нового блока переписки, что исключало влияние предыдущих сообщений на создаваемый текст. Результаты эксперимента показали, что нейронная сеть демонстрирует высокий уровень точности в генерации текстов, соответствующих заданным критериям, особенно в контексте делового стиля.

Из 43 фишинговых писем, 23 были написаны в деловом стиле, а 20 — в дружеском. При этом деловой стиль оказался более успешно имитируемым нейросетью. В 93 % случаев нейронная сеть адекватно поддерживала строгость конструкций, лаконичность и нейтральный тон изложения, что является характерной особенностью деловой переписки. Письма в деловом стиле были правильно структурированы, соблюдали форматирование, включали место для вставки фишинговых ссылок и предлагали шаблоны, которые злоумышленники могли бы дополнить личными данными. Однако в 7 % случаев, что соответствует трём экспериментальным блокам, нейросеть не смогла адекватно поддержать заданный стиль. Эти случаи отличались нарушением формата письма, ошибками в обращениях и добавлением избыточного текста, что снижало правдоподобность создаваемого сообщения.

В отношении писем в дружеском стиле результаты оказались менее удовлетворительными. Нейросеть генерировала тексты, которые, хотя и сохраняли общий дух дружеской переписки, часто демонстрировали признаки гиперболизированности, отсутствия естественного сленга и чрезмерной длины сообщений. Эти паттерны могли насторожить получателей и вызвать подозрения относительно подлинности письма. Злоумышленники, использующие такие письма для фишинга, могли столкнуться с низкой эффективностью атак, что требует дополнительной доработки алгоритмов в этом направлении.

Таким образом, результаты экспериментов показали, что нейронная сеть GPT-3.5 успешно генерирует письма в деловом стиле, что делает её мощным инструментом для создания фишинговых атак в официальной переписке. Однако тексты, сгенерированные в дружеском стиле, остаются менее правдоподобными, что указывает на необходимость дальнейшего совершенствования алгоритмов для более точной имитации естественного общения.

Заключение

Таким образом, результаты исследования демонстрируют, что нейронные сети, такие как GPT-3.5, являются мощным инструментом для создания фишинговых писем, особенно в деловом стиле. Эти модели могут генерировать тексты, которые соответствуют строгим стандартам деловой переписки, что делает их очень эффективными для использования в рамках фишинговых атак. Однако в случае с дружеским стилем возникают трудности в имитации естественного общения, что требует дальнейших усовершенствований в алгоритмах.

Исследование также подчёркивает важность усовершенствования методов защиты от фишинга, включая использование алгоритмов машинного обучения и нейросетевых технологий для обнаружения и предотвращения таких угроз. В условиях быстрого развития технологий искусственного интеллекта, продолжающие исследования в области имитации стилей общения и создания более сложных моделей для фишинговых атак становятся критически важными. Разработка и внедрение эффективных средств защиты, основанных на искусственном интеллекте, будет играть ключевую роль в борьбе с современными киберугрозами.

Таким образом, несмотря на успехи нейросетей в создании правдоподобных фишинговых писем, необходимо продолжать развивать методы защиты, чтобы эффективно противостоять растущим угрозам, связанным с использованием ИИ для преступных целей.

ЛИТЕРАТУРА

1. Терехова Е.С., Пучкова Н.Н., Новикова Л.В. Социальная инженерия в контексте информационной безопасности // Научный журнал «Концепт». — 2024. — Т. 10–11. — С. 1–3. [Электронный ресурс]. Режим доступа: https://cyberleninka.ru/article/n/analiz-vostrebovannosti-ispolzovaniya-neyrosetey-dlya-resheniya-uchebnyh-zadach (дата обращения: 17.04.2025).
2. Демидова-Петрова Е.В., Зотина Е.В. Телефонное мошенничество: современные угрозы и вызовы // Всероссийский криминологический журнал. — 2024. — Т. 18, № 4. — С. 341–348. [Электронный ресурс]. Режим доступа: https://cyberleninka.ru/article/n/telefonnoe-moshennichestvo-sovremennye-ugrozy-i-vyzovy?ysclid=makqhgv3aq945452092 (дата обращения: 10.05.2025).
3. Бардасова И.А., Волкова Е.А. Обнаружение аномалий в электронных письмах с помощью машинного обучения // Международный научный журнал «ВЕСТНИК НАУКИ». — 2024. — № 5 (78), Т. 4. — С. 1351–1354. [Электронный ресурс]. Режим доступа: https://cyberleninka.ru/article/n/obnaruzhenie-anomaliy-v-elektronnyh-pismah-s-pomoschyu-mashinnogo-obucheniya (дата обращения: 10.04.2025).
4. Шишкин С.Р. Имитационное моделирование в сфере защиты информации с применением нейросетей // Журнал «Экономика и качество систем связи». — 2025. — С. 132–136. [Электронный ресурс]. Режим доступа: https://cyberleninka.ru/article/n/imitatsionnoe-modelirovanie-v-sfere-zaschity-informatsii-s-primeneniem-neyrosetey (дата обращения: 12.05.2025).
5. Корнюхина С.П., Лапонина О.Р. Исследование возможностей алгоритмов глубокого обучения для защиты от фишинговых атак // International Journal of Open Information Technologies. — 2023. — Vol. 11. — С. 163–169. [Электронный ресурс]. Режим доступа: https://cyberleninka.ru/article/n/issledovanie-vozmozhnostey-algoritmov-glubokogo-obucheniya-dlya-zaschity-ot-fishingovyh-atak (дата обращения: 17.04.2025).
6. Дурдыев А.Г., Аннасапаров Г.Г., Дурдыев Р.А. История и эволюция информационных технологий // Журнал «Наука и мировоззрение». — 2024. — С. 3–4. [Электронный ресурс]. Режим доступа: https://cyberleninka.ru/article/n/istoriya-i-evolyutsiya-informatsionnyh-tehnologiy/viewer (дата обращения: 17.04.2025).
7. Даниленко Ю.А. Использование искусственного интеллекта в преступных целях: уголовно-правовая характеристика // Ученые записки Крымского федерального университета имени В.И. Вернадского. Юридические науки. — 2023. — Т. 9 (75), № 4. — С. 232–240. [Электронный ресурс]. Режим доступа: https://cyberleninka.ru/article/n/ispolzovanie-iskusstvennogo-intellekta-v-prestupnyh-tselyah-ugolovno-pravovaya-harakteristika-1 (дата обращения: 12.05.2025).
8. Беседина В. Актуальные киберугрозы: III квартал 2024 года // Positive Technologies. — 2024. — 5 ноября. [Электронный ресурс]. Режим доступа: https://ptsecurity.com/ru-ru/research/analytics/aktualnye-kiberugrozy-iii-kvartal-2024-goda/#id3 (дата обращения: 12.05.2025).
9. Тренды фишинговых атак на организации в 2022–2023 годах // Positive Technologies. — 2024. — 14 февраля. [Электронный ресурс]. Режим доступа: https://ptsecurity.com/ru-ru/research/analytics/phishing-attacks-on-organizations-in-2022-2023/ (дата обращения: 12.05.2025).
10. Positive Technologies: больше половины успешных атак с использованием вредоносного ПО начинаются с фишинга // Positive Technologies. — 2024. — 14 мая. [Электронный ресурс]. Режим доступа: https://ptsecurity.com/ru-ru/about/news/positive-technologies-bolshe-poloviny-uspeshnyh-atak-s-ispolzovaniem-vredonosnogo-po-nachinayutsya-s-fishinga/ (дата обращения: 12.05.2025).