ИССЛЕДОВАНИЕ ФИШИНГОВЫХ АТАК, СГЕНЕРИРОВАННЫХ ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ, И МЕТОДОВ ИХ ОБНАРУЖЕНИЯ

Жовлиева Мафтуна Абдуназаровна
Термезский государственный университет,
магистрантка 1-го курса специальности,
«Компьютерные системы и их программное обеспечение»
E-mail: maftunajovliyeva9@gmail.com
Doi: https://doi.org/10.5281/zenodo.20372002

Аннотация. В статье рассматривается влияние генеративного искусственного интеллекта на развитие фишинговых атак. Показано, что крупные языковые модели повышают качество и масштабируемость phishing-кампаний, снижая порог входа для злоумышленников. Проведен анализ современных исследований и официальных threat-intelligence-отчетов за 2021–2026 гг., посвященных AI-generated phishing и методам его обнаружения. На основе обзора предложена мультимодальная архитектура обнаружения, объединяющая NLP-анализ текстов, проверку идентичности отправителя, URL/HTML-анализ и элементы threat-intelligence. Особое внимание уделено безопасной методологии формирования датасета с использованием синтетических примеров и фиктивных доменов. Экспериментальная часть представлена в виде воспроизводимого лабораторного протокола и сравнительного анализа опубликованных результатов. Полученные выводы показывают перспективность гибридных подходов для обнаружения AI-generated phishing.
Ключевые слова: фишинг, социальная инженерия, генеративный ИИ, LLM, phishing detection, URL analysis, кибербезопасность.

SUN’IY INTELLEKT YORDAMIDA YARATILGAN FISHING HUJUMLARI VA ULARNI ANIQLASH USULLARINI TADQIQ QILISH

Annotatsiya. Maqolada generativ sun’iy intellektning fishing hujumlari rivojlanishiga ta’siri ko‘rib chiqiladi. Yirik til modellari (LLM) phishing kampaniyalarining sifati va masshtablanuvchanligini oshirib, kiberjinoyatchilar uchun kirish chegarasini pasaytirishi ko‘rsatib berilgan.

2021–2026 yillarga oid AI-generated phishing va uni aniqlash usullariga bag‘ishlangan zamonaviy ilmiy tadqiqotlar hamda rasmiy threat-intelligence hisobotlari tahlil qilindi. Tahlil asosida matnlarning NLP-analizi, jo‘natuvchi identifikatsiyasini tekshirish, URL/HTML-analiz va threat-intelligence elementlarini birlashtiruvchi multimodal aniqlash arxitekturasi taklif etildi.

Shuningdek, sintetik misollar va soxta domenlardan foydalanish orqali dataset shakllantirishning xavfsiz metodologiyasiga alohida e’tibor qaratilgan. Eksperimental qism qayta tiklanadigan laboratoriya protokoli va e’lon qilingan natijalarning qiyosiy tahlili ko‘rinishida taqdim etilgan. Olingan xulosalar AI-generated phishingni aniqlashda gibrid yondashuvlarning istiqbolli ekanligini ko‘rsatadi.
Kalit so‘zlar: fishing, ijtimoiy muhandislik, generativ sun’iy intellekt, LLM, phishing detection, URL analysis, kiberxavfsizlik.

A STUDY OF AI-GENERATED PHISHING ATTACKS AND THEIR DETECTION METHODS

Abstract. This paper examines the impact of generative artificial intelligence on the evolution of phishing attacks. Large language models are shown to improve the quality and scalability of phishing campaigns while lowering the barrier to entry for attackers. A review of recent studies and official threat-intelligence reports published between 2021 and 2026 is conducted to analyze AI-generated phishing and modern detection approaches. Based on this review, a multimodal detection architecture is proposed that combines NLP-based text analysis, sender-identity verification, URL/HTML analysis, and threat-intelligence enrichment. Particular attention is paid to the safe generation of synthetic datasets using fictional domains and controlled phishing-style samples. The experimental section is presented as a reproducible laboratory protocol together with a comparative analysis of published results. The findings demonstrate the potential of hybrid approaches for detecting AI-generated phishing attacks.

KEYWORDS: phishing, social engineering, generative AI, large language models, phishing detection, URL analysis, cybersecurity.

ВВЕДЕНИЕ. Фишинг остается одной из наиболее распространенных форм киберпреступности и представляет серьезную угрозу для пользователей и организаций. Развитие генеративного искусственного интеллекта и крупных языковых моделей (LLM) существенно изменило характер phishing-атак, повысив качество, масштабируемость и степень персонализации фишинговых сообщений. Современные AI-модели позволяют злоумышленникам автоматически создавать убедительные spear-phishing-сценарии, снижая требования к технической подготовке атакующих.

Согласно отчетам FBI, NCSC и Microsoft, AI-generated phishing становится все более эффективным и трудным для обнаружения традиционными средствами защиты. В отличие от классического фишинга, современные атаки используют не только текстовые манипуляции, но и комбинируют URL-маскировку, подмену доменов, элементы социальной инженерии и автоматизированную генерацию контента.

В связи с этим актуальной задачей становится разработка методов обнаружения AI-generated phishing, способных анализировать не только текст сообщения, но и дополнительные признаки, включая URL-структуру, данные о домене и контекст взаимодействия пользователя.

Целью данной статьи является анализ современных подходов к обнаружению AI-generated phishing, а также разработка безопасной методологии формирования синтетического датасета и архитектуры гибридного детектора на основе NLP- и URL-analysis методов.

ОБЗОР ЛИТЕРАТУРЫ.

Исследования последних лет показывают, что крупные языковые модели способны генерировать highly personalized phishing- и spear-phishing-сообщения, сопоставимые по эффективности с сообщениями, созданными человеком. Экспериментальные исследования демонстрируют, что AI-generated phishing существенно повышает масштабируемость атак и снижает требования к подготовке злоумышленников.

Современные методы обнаружения AI-generated phishing развиваются по нескольким направлениям. Одним из наиболее распространенных подходов являются transformer- и NLP-модели для классификации email-сообщений. Исследования показывают, что модели семейства BERT и RoBERTa обеспечивают высокую точность обнаружения phishing emails, однако их эффективность снижается при использовании HTML-модификаций, смешанных языков и методов obfuscation.

Отдельное направление связано с использованием LLM-based detectors и explainable warning systems. Такие системы позволяют не только классифицировать подозрительные сообщения, но и формировать объяснения для пользователя, повышая эффективность принятия решений. Также активно развиваются reference-based и multimodal approaches, сочетающие анализ текста, URL-структуры, HTML-контента и доменной информации.

Таким образом, современные исследования показывают, что наиболее перспективными являются гибридные подходы, объединяющие NLP-анализ, URL/HTML-проверку и contextual verification. При этом проблема обнаружения AI-generated phishing остается актуальной из-за высокой адаптивности атак и постоянного развития генеративных моделей.

Таблица 1.

Сравнение traditional phishing и AI-generated phishing

Характеристика	Traditional phishing	AI-generated phishing
Персонализация	Ограниченная	Высокая
Масштабируемость	Средняя	Высокая
Грамматические ошибки	Часто присутствуют	Минимальны
Скорость генерации	Низкая	Высокая
Использование LLM	Нет	Да

Таблица 2.

Основные подходы к обнаружению AI-generated phishing

Подход	*Объект анализа*	*Преимущества*	*Ограничения*
Transformer/NLP models	Email text	Высокая точность классификации	Чувствительность к obfuscation и HTML
LLM-based detection	Email text	Семантический анализ и explainable warnings	Высокие вычислительные затраты
URL/HTML analysis	URL и web-content	Обнаружение phishing webpages	Требуется качественный parsing
Multimodal approaches	Text + URL + HTML	Более высокая устойчивость к evolving attacks	Сложность реализации

МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

В работе предлагается методология формирования датасета для исследования AI-generated phishing, включающая четыре класса данных: legitimate emails, ordinary phishing, AI-generated legitimate messages и AI-generated phishing. Такое разделение позволяет обучать модель выявлять признаки фишингового поведения, а не только различать тексты, созданные человеком или языковой моделью.

Для формирования корпуса используются как открытые phishing- и ham-датасеты, так и синтетические примеры, созданные с помощью LLM в контролируемой среде. В качестве источников legitimate и phishing emails могут использоваться публичные наборы Enron, SpamAssassin, PhishTank и OpenPhish. AI-generated phishing-сообщения

генерируются с применением безопасных шаблонов без использования реальных брендов, доменов и учетных данных.

Основное внимание уделяется email- и URL-анализу. Для ordinary phishing используются обезвреженные текстовые и URL-образцы без активных ссылок и вредоносного содержимого. Синтетические phishing-примеры формируются на основе типовых сценариев социальной инженерии, включая urgency, account verification и support requests.

Для обеспечения безопасности исследования все синтетические сообщения создаются исключительно в лабораторной среде. Используются фиктивные домены формата example.* и отключенные HTML-формы. Разделение train, validation и test выполняется по шаблонам и phishing-кампаниям, что позволяет снизить риск утечки признаков между выборками и избежать искусственного завышения метрик.

Для оценки эффективности моделей используются стандартные метрики классификации: Accuracy, Precision, Recall, F1-score и ROC-AUC.

Основные метрики классификации:

Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision:

$$Precision = \frac{TP}{TP+FP}$$

Recall:

$$Recall = \frac{TP}{TP+FN}$$

F1-score:

$$F1 = 2 \times \frac{Precision * Recall}{Precision+Recall}$$

ROC-AUC:

ROC-AUC используется для оценки качества модели при разных порогах классификации и показывает способность модели отличать phishing-сообщения от legitimate-сообщений

ПРИМЕР ВЫЧИСЛЕНИЯ МЕТРИК

Для оценки качества модели используются стандартные метрики бинарной классификации. Пусть в результате тестирования модели были получены следующие значения:

- TP = 920
- TN = 870
- FP = 80
- FN = 130

Тогда:

Вычисление Accuracy

Формула:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Подставим значения:

$$Accuracy = \frac{920+870}{920+870+80+130}$$

$$Accuracy = \frac{1790}{2000}$$

$$Accuracy = 0.895$$

$$Accuracy = 89.5\%$$

Вычисление Precision

Формула:

$$Precision = \frac{TP}{TP+FP}$$

Подставим значения:

$$Precision = \frac{920}{920+80}$$

$$Precision = \frac{920}{1000}$$

$$Precision = 0.92$$

$$Precision = 92\%$$

Вычисление Recall

Формула:

$$Recall = \frac{TP}{TP+FN}$$

Подставим значения:

$$Recall = \frac{920}{920+130}$$

$$Recall = \frac{920}{1050}$$

$$Recall \approx 0.876$$

$$Recall \approx 87.6\%$$

Вычисление F1-score

Формула:

$$F1 = 2 \times \frac{Precision * Recall}{Precision+Recall}$$

Подставим значения:

$$F1 = \frac{2 \cdot 0{,}92 \cdot 0{,}876}{0{,}92+,0876}$$

$$F1 = \frac{1{,}61184}{1{,}796}$$

$$F1 \approx 0{,}897$$

$$F1 \approx 89.7\%$$

Полученные значения демонстрируют высокую эффективность модели при обнаружении phishing-сообщений.

Таблица 3.

Признаки phishing-сообщений

Признак	*Описание*
Urgency language	Использование срочных требований
Suspicious URLs	Подозрительные ссылки
Domain mismatch	Несоответствие домена
Credential request	Запрос учетных данных
Impersonation	Имитация бренда или отправителя

Таблица 4.

Структура экспериментального датасета

Тип данных	*Класс*	*Источник*
Email	Legitimate	Enron, SpamAssassin
Email	Ordinary phishing	PhishTank, OpenPhish
Email	AI-generated phishing	Синтетические LLM-generated samples
URL/HTML	Benign / phishing	Публичные web snapshots

ПРЕДЛАГАЕМАЯ СИСТЕМА ОБНАРУЖЕНИЯ

Предлагаемая система обнаружения AI-generated phishing основана на мультимодальной архитектуре, объединяющей анализ текста, URL-структуры и HTML-контента. Основная идея системы заключается в том, что современные AI-generated phishing-атаки часто не содержат очевидных лингвистических ошибок, поэтому их обнаружение требует анализа нескольких источников признаков одновременно.

Система включает три основных модуля: 1. NLP-модуль для анализа текста email-сообщений; 2. модуль URL/HTML-анализа; 3. модуль проверки соответствия домена и заявленного отправителя.

Текстовый модуль использует transformer-модели семейства BERT/RoBERTa для классификации phishing-сообщений и выявления признаков социальной инженерии. Дополнительно применяется семантический анализ содержимого письма для определения подозрительных сценариев, связанных с urgency, account verification и impersonation.

URL/HTML-модуль выполняет анализ структуры ссылок, доменных признаков и HTML-контента веб-страниц. Особое внимание уделяется проверке соответствия между заявленным брендом, адресом отправителя и доменом ссылки. Такой подход позволяет повысить устойчивость системы к современным phishing-атакам, использующим подмену доменов и маскировку web-интерфейсов.

Результаты отдельных модулей объединяются с помощью fusion-модели, формирующей итоговую оценку риска phishing-атаки. В зависимости от уровня риска сообщение классифицируется как benign, suspicious или malicious. Дополнительно система может формировать explainable warnings для пользователя.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Экспериментальная часть работы построена как воспроизводимый лабораторный протокол для оценки методов обнаружения AI-generated phishing. Основное внимание уделяется анализу email-сообщений и URL/HTML-признаков с использованием как реальных phishing-датасетов, так и синтетических AI-generated samples.

Для оценки моделей используется стандартное разделение данных на train, validation и test выборки в пропорции 70/15/15, а также 5-fold cross-validation. Разделение выполняется по phishing-кампаниям, шаблонам и доменам, что позволяет снизить риск утечки признаков между выборками и обеспечить более реалистичную оценку качества моделей.

В качестве основных метрик используются Accuracy, Precision, Recall, F1-score и ROC-AUC. Дополнительно оцениваются устойчивость моделей к URL-obfuscation и модификациям phishing-контента.

Сравнительный анализ современных исследований показывает высокую эффективность transformer- и hybrid-based approaches для phishing detection. Модели семейства RoBERTa демонстрируют высокие показатели accuracy при классификации email-сообщений, а multimodal systems обеспечивают более устойчивое обнаружение phishing webpages и AI-generated phishing attacks.

Таблица 3.
Сравнение современных подходов к phishing detection

Работа	*Подход*	*Основной результат*
Heiding et al., 2024	AI-generated spear phishing	Высокая эффективность AI-generated phishing

Работа	*Подход*	*Основной результат*
APOLLO, 2025	LLM-based detection	Высокая точность и explainable warnings
KnowPhish, 2024	Multimodal phishing detection	Устойчивость к phishing webpages
Meléndez et al., 2024	RoBERTa-based detection	Высокая accuracy email classification
PiMRef, 2025	Reference-based verification	Эффективная проверка sender/domain consistency

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Одной из основных проблем обнаружения AI-generated phishing является высокая адаптивность современных атак. Злоумышленники могут изменять стиль сообщений, использовать paraphrasing, маскировку URL и комбинировать элементы социальной инженерии, что снижает эффективность традиционных text-only detectors.

Проведенный анализ подтверждает, что AI-generated phishing представляет собой реальную угрозу для пользователей и организаций. Использование генеративных моделей позволяет автоматизировать создание highly personalized phishing messages и повышает масштабируемость атак.

Практическое внедрение систем обнаружения AI-generated phishing должно основываться на комплексном анализе текста, URL-структуры и доменной информации. Дополнительно перспективным направлением является использование explainable warnings, повышающих информированность пользователя и качество принятия решений.

Заключение. В статье рассмотрено влияние генеративного искусственного интеллекта на развитие phishing-атак и проанализированы современные подходы к их обнаружению. Показано, что AI-generated phishing отличается высокой степенью автоматизации и персонализации, что усложняет его обнаружение традиционными средствами защиты.

На основе анализа современных исследований предложена архитектура гибридной системы обнаружения, объединяющая NLP-анализ текстов, URL/HTML-проверку и contextual verification. Также предложена безопасная методология формирования синтетического датасета для исследования AI-generated phishing в лабораторной среде.

Полученные результаты показывают перспективность мультимодальных и hybrid-based approaches для обнаружения AI-generated phishing attacks. В качестве дальнейшего направления исследований рассматривается разработка и экспериментальная оценка собственных phishing detection models на основе современных transformer-архитектур.

СПИСОК ЛИТЕРАТУРЫ

1. F. Heiding et al., “Devising and Detecting Phishing Emails Using Large Language Models,” IEEE Access, 2024.
2. M. Bethany et al., “Large Language Model Lateral Spear Phishing: A Comparative Study in Large-Scale Organizational Settings,” arXiv, 2024.
3. F. Heiding et al., “Evaluating Large Language Models’ Capability to Launch Fully Automated Spear Phishing Campaigns,” arXiv, 2024.

4. M. Weinz et al., “The Impact of Emerging Phishing Threats,” ASIA CCS, 2025.
5. G. Desolda et al., “APOLLO: A GPT-based Tool to Detect Phishing Emails,” Proc. ACM Hum.-Comput. Interact., 2025.
6. Y. Li et al., “KnowPhish: Large Language Models Meet Multimodal Knowledge Graphs,” USENIX Security, 2024.
7. R. Liu et al., “PiMRef: Detecting and Explaining Ever-evolving Spear Phishing Emails,” arXiv, 2025.
8. R. Meléndez et al., “Comparative Investigation of Transformer Models for Phishing Email Detection,” Electronics, 2024.
9. M. Ibrahim and R. Elhafiz, “Phishing Email Detection Using BERT and RoBERTa,” Computation, 2026.
10. H. Jiang et al., “D-PhishNet: URL and HTML Feature Fusion in Phishing Detection,” Computer Networks, 2025.