УДК 004.056.5

С.В. Корякин, srgkoryakin1@gmail.com
Институт информационных технологий КГТУ им. И.Раззакова
Институт машиноведения автоматики и геомеханики НАН КР
В.В. Ким, vyacheslav_kim.2003@mail.ru
Институт информационных технологий КГТУ им. И. Раззакова

ОБЗОР МЕТОДОВ РАЗРАБОТКИ ПОДСИСТЕМ ЗАЩИТЫ ОТ ФИШИНГОВЫХ АТАК С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

Данная работа посвящена анализу методов разработки подсистемы SOC для защиты от фишинговых атак с использованием технологий искусственного интеллекта (ИИ) и машинного обучения (МО). Рассмотрены основные методы обнаружения фишинга, проведен обзор коммерческих систем защиты и их функциональных возможностей. Особое внимание уделено сравнению моделей ИИ для выявления фишинговых атак с оценкой их точности, скорости работы и адаптивности. Также приведены примеры применения ИИ в действующих системах безопасности и рассмотрены возможные сложности внедрения таких решений, включая настройку и риски ложных срабатываний. В заключение даны рекомендации по выбору подходящих технологий и инструментов для разработки подсистемы SOC, позволяющей эффективно выявлять и блокировать фишинговые атаки с применением ИИ.

Ключевые слова: фишинг, ИИ, машинное обучение, SOC, кибербезопасность, нейронные сети, анализ текста, автоматизация, деревья решений, случайные леса, ансамблевые методы, машины опорных векторов (SVM), логистическая регрессия, градиентный бустинг, K-means, BERT

Введение

Одной из наиболее распространенных киберугроз являются фишинговые атаки, нацеленные на компрометацию корпоративных данных. Более 80% кибератак в 2024 году начинались с фишинга. Традиционные методы защиты, такие как черные списки и фильтры ключевых слов, не противостоят новым, сложным угрозам. В этой статье рассматриваются преимущества использования технологий искусственного интеллекта (ИИ) и машинного обучения (МО) в защите от фишинговых атак, сравнение их с традиционными методами и перспективы развития. [1]

В 2024 году наблюдался значительный рост фишинговых атак на онлайн-магазины и финансовые компании. Согласно отчету «Лаборатории Касперского», количество попыток перехода по фишинговым ссылкам увеличилось на 26% по сравнению с предыдущим годом, достигнув 893 216 170 случаев. [1]

Согласно исследованию, среднее количество фишинговых угроз на каждый бренд выросло с 495 до 734. Ежедневно злоумышленники создают около четырех новых фишинговых ресурсов под каждый целевой бренд. Подавляющее большинство этих сайтов (94%) нацелено на кражу данных банковских карт, оставшиеся 6% предназначены для хищения учетных данных пользователей.

Исследователи выявили следующее распределение вредоносных ресурсов по типам:
— 70% — поддельные веб-сайты
— 13% — каналы и боты в мессенджерах
— 11% — группы в социальных сетях
— 4% — мобильные приложения
— 2% — объявления на торговых площадках.

Сфера электронной коммерции также находится под серьезным давлением киберпреступников. Количество фишинговых ресурсов в этом секторе увеличилось на 33,7%,

а среднее число угроз на один бренд достигло 405. Особую озабоченность вызывает рост числа скам-ресурсов на 20,7%, достигший показателя 1333 угрозы на бренд. [2]

В 2024 году в странах СНГ, включая Кыргызстан, наблюдался рост кибератак, в том числе фишинговых. Злоумышленники активно использовали методы социальной инженерии и вредоносное программное обеспечение для компрометации учетных записей сотрудников различных организаций. Особенно уязвимыми были сотрудники, работающие с конфиденциальной информацией, такие как специалисты IT, HR и топ-менеджмент. [3]

Фишинг остается одной из самых распространенных киберугроз, и, согласно статистике, количество таких атак продолжает расти. Современные фишинговые атаки становятся все сложнее, активно применяются технологии искусственного интеллекта (ИИ), что делает традиционные методы защиты менее эффективными.

В связи с этим возрастает необходимость использования ИИ и машинного обучения для защиты от фишинга. Такие технологии позволяют быстрее выявлять новые угрозы, снижать количество ложных срабатываний, автоматизировать обработку инцидентов и сокращать нагрузку на специалистов SOC. Это повышает эффективность защиты и снижает затраты на ее поддержку.

Методы обнаружения фишинга

С развитием технологий фишинговые атаки становятся все более сложными, требуя усовершенствованных методов их обнаружения. Наиболее перспективным направлением в борьбе с фишингом является использование искусственного интеллекта (ИИ), который позволяет анализировать атаки на более глубоком уровне и адаптироваться к новым угрозам. Однако традиционные методы также находят применение, особенно в комбинации с ИИ-алгоритмами.

Традиционные методы обнаружения фишинга. Исторически фишинг выявляли с помощью правил и сигнатур:

- Методы на основе правил используют заранее заданные эвристики, например, проверку подозрительных URL или подозрительных заголовков писем. [4]
- Сигнатурный анализ опирается на базы данных известных фишинговых шаблонов.

Преимущества традиционных методов:

- Простота реализации.
- Высокая скорость обнаружения известных угроз.

Недостатки:

- Неэффективность против новых, неизвестных атак.
- Требуется постоянное обновление баз данных.
- Высокая вероятность ложных срабатываний [4].

В современных условиях эти методы слабо защищают от фишинга, но могут дополнять другие подходы, например, ИИ-аналитику. Общие способы обнаружения описаны на рисунке 1 [5]

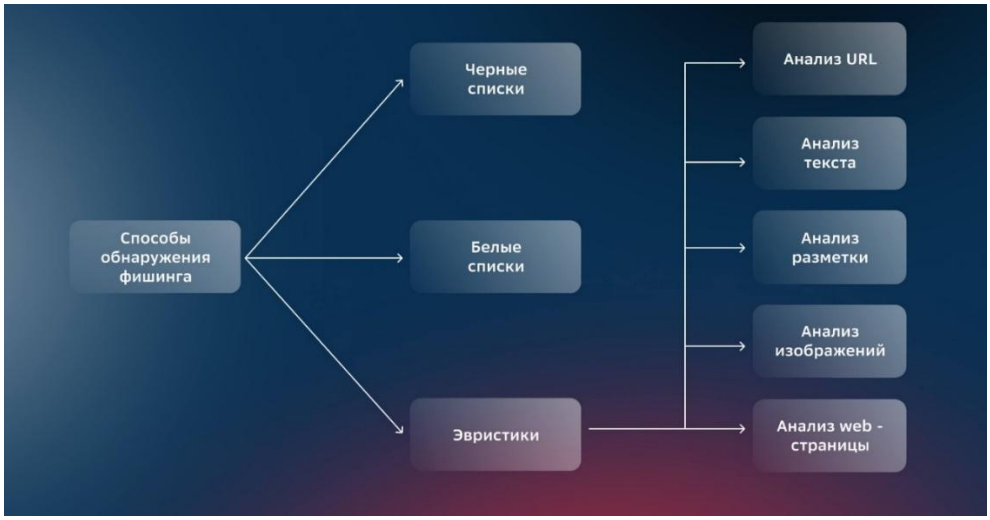

Рисунок 1 – Традиционные методы обнаружения фишинга

Методы машинного обучения

Алгоритмы машинного обучения (МО) анализируют массивы данных, выявляя фишинг без необходимости предварительно заданных правил. [4]

Контролируемое обучение (Supervised Learning)

Требует размеченного набора данных с примерами фишинговых и безопасных писем. [5]

Популярные алгоритмы:

- Деревья решений – легко интерпретируемые модели, анализирующие признаки писем.
- Случайный лес – повышает точность за счет комбинации нескольких деревьев решений.
- SVM (метод опорных векторов) – эффективно классифицирует сложные случаи. [5]

Неконтролируемое обучение (Unsupervised Learning)

Используется, когда нет размеченных данных. Позволяет находить новые виды атак:

- Кластеризация (k-means, DBSCAN) – группирует похожие письма и выделяет аномалии.
- Обнаружение аномалий (Isolation Forests) – находит отклонения от нормального поведения. [4]

Преимущества МО:

- Высокая точность при достаточном количестве данных.
- Способность выявлять ранее неизвестные атаки.
- Автоматическая адаптация к новым угрозам [4].

Недостатки:

- Требует больших объемов данных для обучения.
- Возможны ложные срабатывания [5].

Обработка естественного языка (NLP) для обнаружения фишинга

ИИ активно используется для анализа текстов писем с помощью методов обработки естественного языка (NLP). Это позволяет выявлять фишинговые атаки на основе содержания. [4]

Методы NLP включают:

- Классификацию текста (определяет, является ли письмо фишингом).
- Анализ тональности (подозрительные обращения, срочность, запугивание).
- Семантический анализ (поиск лексических признаков фишинга).

Современные модели NLP (например, BERT, GPT) анализируют текст глубже, чем простые эвристики.

Преимущества NLP:

- Учитывает смысл текста, а не только формальные признаки.
- Выявляет сложные фишинговые атаки [4].

Недостатки:

- Требует значительных вычислительных ресурсов.
- Сложность интерпретации [5].

Глубокое обучение в обнаружении фишинга

Глубокое обучение позволяет обнаруживать фишинг на более сложном уровне, автоматически выявляя закономерности в данных. [5]

Основные модели:

- CNN (сверточные нейронные сети) – анализируют URL и структуру письма.
- RNN и LSTM (рекуррентные сети) – обрабатывают текстовые последовательности, выявляя подозрительные фразы.
- Трансформеры (BERT, GPT) – отлично анализируют контекст, выявляя даже замаскированные атаки.

Преимущества:

- Высокая точность при анализе сложных атак.
- Способность учитывать контекст сообщений [5].

Недостатки:

- Высокие вычислительные затраты.
- Необходимость в больших объемах обучающих данных.

Общие различия машинного обучения и глубокого обучения изображены на рисунке 2.

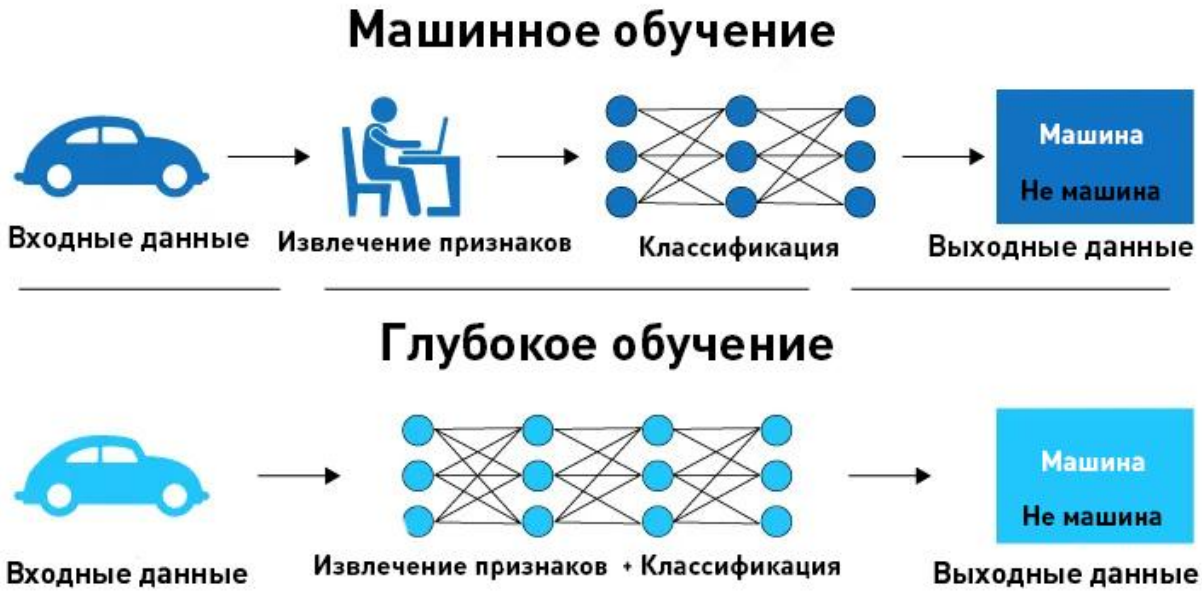

Рисунок 2 –Машинное и глубокое обучение

В таблице 1 описано сравнение различных методов защиты от фишинга по принципу работы, их преимущества, недостатки и актуальность.

Таблица 1 – Сравнение методов защиты от фишинга

Метод	Принцип работы	Преимущества	Недостатки	Актуальность

Методы на основе правил	Проверка заголовков, URL, ключевых слов по заданным эвристикам	Простота и скорость работы. Не требует сложных вычислений	Уязвим для новых атак. Высокая доля ложных срабатываний	Устарел, может использоваться в сочетании с ИИ
Сигнатурный анализ	Сравнение с базами известных фишинговых шаблонов	Быстродействие. Высокая точность для известных атак	Не выявляет новые угрозы. Требует постоянного обновления баз	Эффективен только в комбинации с другими методами
Машинное обучение (ML)	Обучение модели на размеченных данных	Высокая точность. Адаптация к новым угрозам	Требует обучающего набора данных. Возможны ложные срабатывания	Эффективен, но требует качественных данных
Обработка естественного языка (NLP)	Анализ текста, структуры писем, тональности	Учитывает контекст. Выявляет сложные атаки	Высокая нагрузка на вычислительные мощности	Перспективный метод, особенно с трансформерами (BERT, GPT)
Глубокое обучение (DL)	Автоматический анализ признаков с помощью нейросетей (CNN, RNN, трансформеры)	Высокая точность. Выявляет сложные фишинговые схемы	Требует больших данных и ресурсов. Сложен в интерпретации	Один из наиболее эффективных подходов, но требует значительных вычислительных мощностей

ИИ-методы, особенно ML, NLP и DL, обеспечивают высокую точность и адаптацию к новым фишинговым атакам, превосходя традиционные подходы. Хотя они требуют качественных данных и вычислительных ресурсов, их применение является наиболее перспективным для эффективного обнаружения и предотвращения фишинга.

Анализ коммерческих систем обнаружения фишинга.

PhishLabs. PhishLabs предлагает комплексное решение для защиты от фишинговых атак. Система использует машинное обучение, анализ ссылок и URL-адресов, а также предоставляет детектирование фишинговых сайтов и фишинговых писем.

Преимущества:

- Полный анализ угроз в реальном времени.
- Поддержка различных типов фишинга, включая фишинг через социальные сети.
- Оповещения и отчеты для управления инцидентами.

Недостатки:

- Высокая стоимость, особенно для малого и среднего бизнеса.
- Требуется интеграция с другими системами безопасности.

Цена: от $3,000 в год, в зависимости от масштаба и потребностей организации. [6]

Cofense (ранее PhishMe)

Cofense предлагает решения для обнаружения фишинга и обучения сотрудников, чтобы предотвратить атаки через фишинговые письма. Система использует как ручные, так и автоматические методы для обнаружения фишинговых угроз.

Преимущества:

- Высокая точность в классификации фишинговых писем.
- Инструменты для обучения сотрудников на основе реальных атак.
- Обширная база данных фишинговых сообщений.

Недостатки:

- Требует постоянного обновления данных об угрозах.
- Дорогое решение для небольших компаний.

Цена: Цены варьируются от $2,000 до $10,000 в год, в зависимости от функционала и количества пользователей. [7]

Barracuda PhishLine

Barracuda PhishLine предлагает комплексное решение для обучения сотрудников и обнаружения фишинговых атак. Система использует как обучение через имитацию фишинга, так и активную защиту от фишинговых писем и ссылок.

Преимущества:

- Инструменты для создания тестов фишинговых атак и обучения сотрудников.
- Поддержка различных каналов (электронная почта, социальные сети).
- Интеграция с другими продуктами Barracuda для комплексной защиты.

Недостатки:

- Ограниченные возможности для обнаружения фишинга на уровне веб-сайтов.
- Не всегда эффективно работает против продвинутых фишинговых техник.

Цена: от $1,200 до $10,000 в год в зависимости от выбранных опций. [8]

Proofpoint. Proofpoint — это система, предназначенная для защиты от фишинговых атак через электронную почту и веб-сайты. Использует технологии ИИ для анализа письма и обнаружения подозрительных ссылок и вложений.

Преимущества:

- Высокая точность в фильтрации фишинговых писем.
- Обширный анализ угроз на основе ИИ и машинного обучения.
- Хорошая интеграция с другими корпоративными инструментами безопасности.

Недостатки:

- Высокая стоимость для малых и средних предприятий.
- Не всегда подходит для защиты от фишинга через каналы, отличные от электронной почты.

Цена: Цены начинаются от $15,000 в год для малого бизнеса и могут достигать $100,000+ для крупных организаций. [9]

KnowBe4

KnowBe4 предлагает систему для обучения сотрудников и обнаружения фишинговых атак. Включает в себя инструменты для имитации фишинга, обучение персонала и защиты от фишинговых угроз.

Преимущества:

- Мощный инструмент для тестирования сотрудников на фишинг.
- Постоянные обновления базы данных фишинговых угроз.
- Хорошая интеграция с системами безопасности.

Недостатки:

- Ограниченная способность в обнаружении фишинга на уровне веб-сайтов.
- Некоторые пользователи сообщают о сложности в интерфейсе и настройках.

Цена: от $1,000 до $15,000 в год в зависимости от пакета и количества пользователей. [10]

Mimecast

Mimecast предоставляет облачную защиту от фишинговых атак, включая фильтрацию электронной почты, защиту от вредоносных ссылок и обучение пользователей.
Преимущества:

- Простой интерфейс и хорошая настройка под организацию.
- Высокая точность в обнаружении фишинговых писем и вложений.
- Интеграция с другими решениями для безопасности в облаке.

Недостатки:

- Высокая цена на более высокие уровни защиты.
- Ограниченные возможности защиты от фишинга через социальные сети.

Цена: от $5,000 в год для малого бизнеса до $50,000+ для крупных организаций. [11]

В таблице 2 описано сравнение вышеупомянутых коммерческих систем обнаружения фишинга, а также разработка своей системы.

Таблица 2 – Сравнение систем

Система	Описание	Преимущества	Недостатки	Цена
PhishLabs	Комплексное решение для защиты от фишинга с анализом ссылок и URL	Полный анализ угроз в реальном времени; поддержка разных типов фишинга; Оповещения	Высокая стоимость; требуется интеграция с другими системами	От $3,000 в год
Cofense (PhishMe)	Обнаружение фишинговых писем и обучение сотрудников	Высокая точность классификации; Инструменты обучения	Требует постоянного обновления данных; дорогое для небольших компаний	От $2,000 до $10,000 в год
Barracuda PhishLine	Имитация фишинга и обучение сотрудников	Создание тестов фишинговых атак; поддержка разных каналов	Ограниченные возможности обнаружения фишинга на веб-сайтах	От $1,200 до $10,000 в год
Proofpoint	Защита от фишинга через email и веб-сайты, ИИ-анализ	Высокая точность фильтрации; обширный анализ угроз	Высокая стоимость; ограниченная защита вне email	От $15,000 до $100,000+ в год
KnowBe4	Обучение сотрудников, имитация фишинга	Мощные инструменты тестирования; частые обновления базы угроз	Ограниченная защита веб-сайтов; сложный интерфейс	От $1,000 до $15,000 в год
Mimecast	Облачная защита email, фильтрация вредоносных ссылок	Простой интерфейс; интеграция с облачными сервисами	Высокая цена; ограниченная защита соцсетей	От $5,000 до $50,000+ в год
Разработка собственной системы	Использование модели ИИ для анализа текстов на фишинг	Гибкость настройки под бизнес; более низкая стоимость эксплуатации	Требует начальных затрат на разработку и обучение	Разовые затраты на разработку, дальнейшее обслуживание дешевле

Разработка собственной ИИ-системы требует начальных вложений, но в долгосрочной перспективе снижает затраты и позволяет гибко адаптировать защиту под бизнес-процессы. Коммерческие решения точны, но дороги и ограничены, тогда как своя система дает независимость и расширяемость.

Сравнение моделей ИИ для создания системы обнаружения фишинга

Для создания собственной системы необходимо выбрать модель ИИ, которая будет удовлетворять всем требованиям системы. Для этого нужно сравнить известные модели ИИ.

В таблице 3 приведено сравнение различных моделей ИИ для создания своей системы обнаружения фишинга.

Таблица 3 – Сравнение моделей ИИ

Модель	Описание	Плюсы	Минусы
BERT[12]	Трансформер с двусторонним контекстом	Высокая точность, отличная работа с контекстом	Высокие вычислительные требования, много данных
RoBERTa[13]	Улучшенная версия BERT	Повышенная точность, лучше для длинных текстов	Высокие затраты, сложная интерпретация
DistilBERT[14]	Облегченная версия BERT	Почти такая же точность, меньше ресурсов	Чуть хуже на сложных паттернах
FastText[19]	Быстрая модель для коротких текстов	Высокая скорость, низкие требования	Не учитывает контекст, неэффективен против сложных атак
XLNet[15]	Улучшенная версия трансформера	Высокая точность, лучше для сложных текстов	Высокие затраты, сложность настройки
GPT-4[17]	Генеративная модель	Отлично анализирует сложные схемы	Требует много ресурсов, трудно объяснить решения
InstructGPT[18]	GPT-3 для инструкций	Точные предсказания, анализ контекста	Требует подстройки, высокие затраты
T5	Модель для текстовых задач	Гибкость, адаптивность	Высокие ресурсы, долгое обучение
ERNIE[20]	Китайская версия BERT	Улучшенное понимание текста	Ограниченные ресурсы и документация
LightGBM[21]	Градиентный бустинг для структурированных данных	Быстро обучается, хорошо для табличных данных	Плохо с необработанным текстом
Hybrid Model (BERT + LightGBM)	BERT анализирует текст, LightGBM предсказывает	Сочетание точности и скорости	Сложность настройки, большие ресурсы

Графики сравнения моделей

Также при выборе модели ИИ необходимо обращать внимание на процент ложных срабатываний, чтобы он был не сильно высоким.

Уровень ложных срабатываний (%) для разных моделей описан в таблице 4.

Таблица 4 – Уровень ложных срабатываний

Модель	Ложные положительные (FP), %	Ложные отрицательные (FN), %	Общий процент
BERT	3.5%	4.2%	7.2%
RoBERTa	2.8%	3.5%	6.3%
DistilBERT	4.0%	5.1%	9.1%
FastText	7.5%	9.2%	16.6%
XLNet	2.7%	3.3%	6%
GPT-4	2.1%	3.0%	5.1%
InstructGPT	2.3%	3.5%	5.8%
T5	3.1%	4.0%	7.1%
ERNIE	2.9%	3.7%	6.6%
LightGBM	6.5%	8.0%	14.5%
Hybrid (BERT + LightGBM)	2.9%	3.5%	6.4%

На рисунке 3 изображен график, демонстрирующий процент ложных срабатываний у различных моделей ИИ.

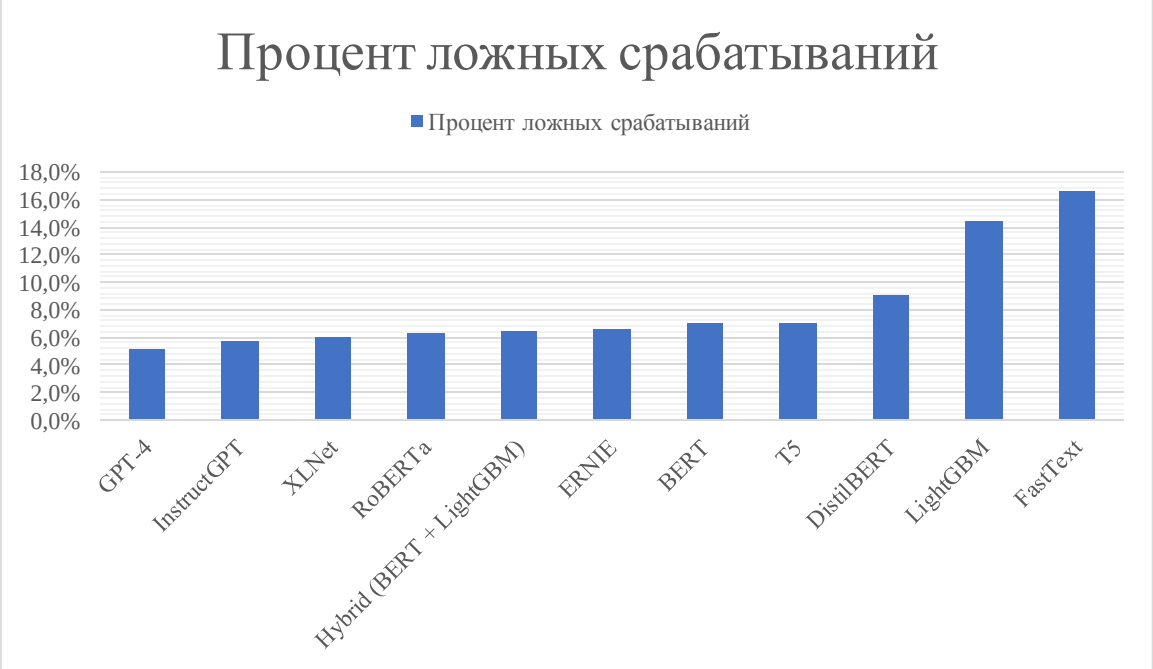

Рисунок 3 – График, демонстрирующий процент ложных срабатываний

Среднее время обработки одного запроса (мс) – еще один параметр, который влияет на скорость работы подсистемы – время обработки запросов. Чем быстрее идет обработка, тем больше писем подсистема сможет анализировать и обрабатывать.

В таблице 5 приведено среднее время обработки запроса для каждой модели ИИ.

Таблица 5 – Среднее время обработки одного запроса

Модель	Среднее время обработки (мс)
BERT	280
RoBERTa	320
DistilBERT	150
FastText	10
XLNet	350
GPT-4	450
InstructGPT	420
T5	380
ERNIE	300
LightGBM	15
Hybrid (BERT + LightGBM)	140

На графике сопоставлено среднее время обработки запроса различных моделей ИИ в порядке увеличения времени.

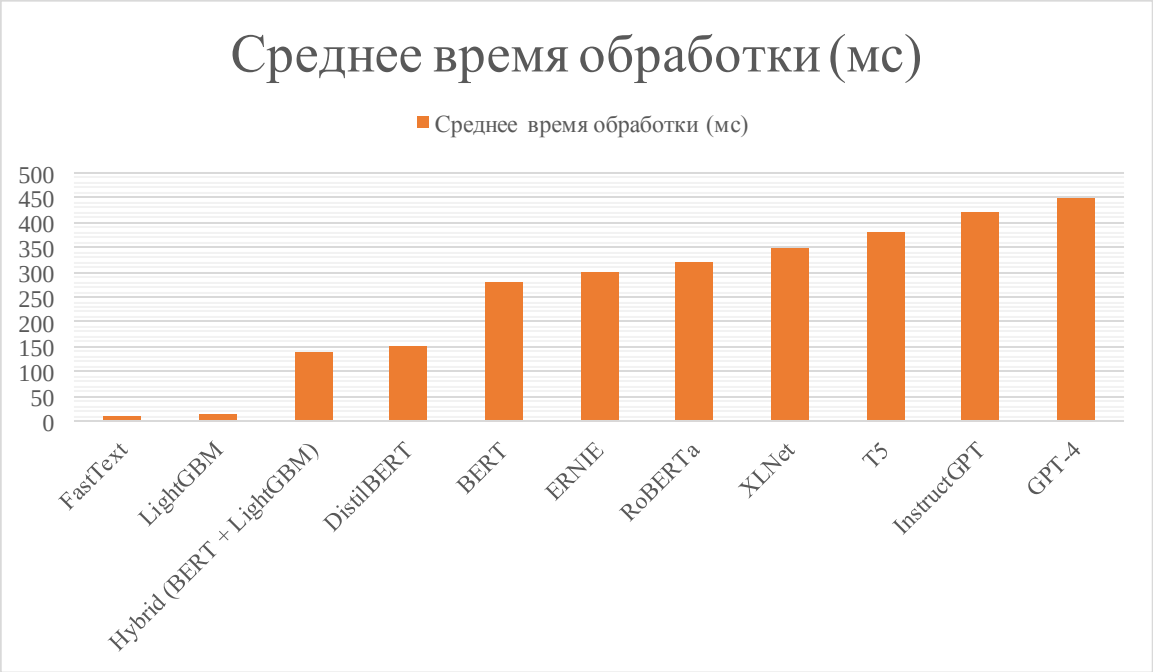

Рисунок 4 – График, демонстрирующий среднее время обработки одного запроса

Для максимальной точности: лучше использовать RoBERTa, GPT-4 или XLNet.

Для баланса скорости и точности: подойдут BERT, DistilBERT, ERNIE или гибридные модели.

Для быстрого, но менее точного анализа: можно использовать FastText или LightGBM.

Оптимальный вариант: гибрид BERT + LightGBM, так как он сочетает хорошую точность и приемлемую скорость.

Реальные примеры применения ИИ в борьбе с фишингом. Google Safe Browsing

Google использует искусственный интеллект для анализа веб-сайтов и выявления фишинговых страниц. Система анализирует сайты на наличие подозрительных ссылок, а также проверяет их репутацию, применяя различные алгоритмы машинного обучения. По данным Google, использование таких технологий помогло значительно снизить количество успешных фишинговых атак. [27] Anti-Phishing Working Group (APWG). APWG использует ИИ для сбора и анализа данных о фишинговых сайтах и письмах, делая их доступными для

общественности. Системы на основе ИИ автоматизируют процесс идентификации фишинговых атак, что помогает быстрее реагировать на угрозы. [27]

Проведение исследования Kaspersky

Kaspersky проводил исследование, насколько хорошо ChatGPT может обнаруживать фишинг. По итогам нескольких проверок, если дать ChatGPT только URL-ссылку на веб-сайт, то результат после всех ошибок и отказов будет следующий: «Мы остались с набором данных из 2322 фишинговых и 2943 легитимных URL-адресов. Итоговые показатели:

- Доля обнаруженного фишинга (detection rate, DR): **87,2%.**
- Доля ложных срабатываний (false positive rate, FPR): **23,2%».**

Также есть статья «URLNet: изучение представления URL с помощью глубокого обучения для обнаружения вредоносных URL». В данной статье предоставляется ROC-кривая, иллюстрирующая достигаемые доли истинных обнаружений (уровень обнаружения) и ложных.

При этом ChatGPT хорошо извлекает название компаний из ссылок, указывает данные на сервис WhoIs, на контент веб-сайта, на срок действия сертификата SSL и т.п.

На рисунке 6 изображена ROC-кривая, иллюстрирующая достигаемые доли истинных обнаружений (уровень обнаружения) и ложных.

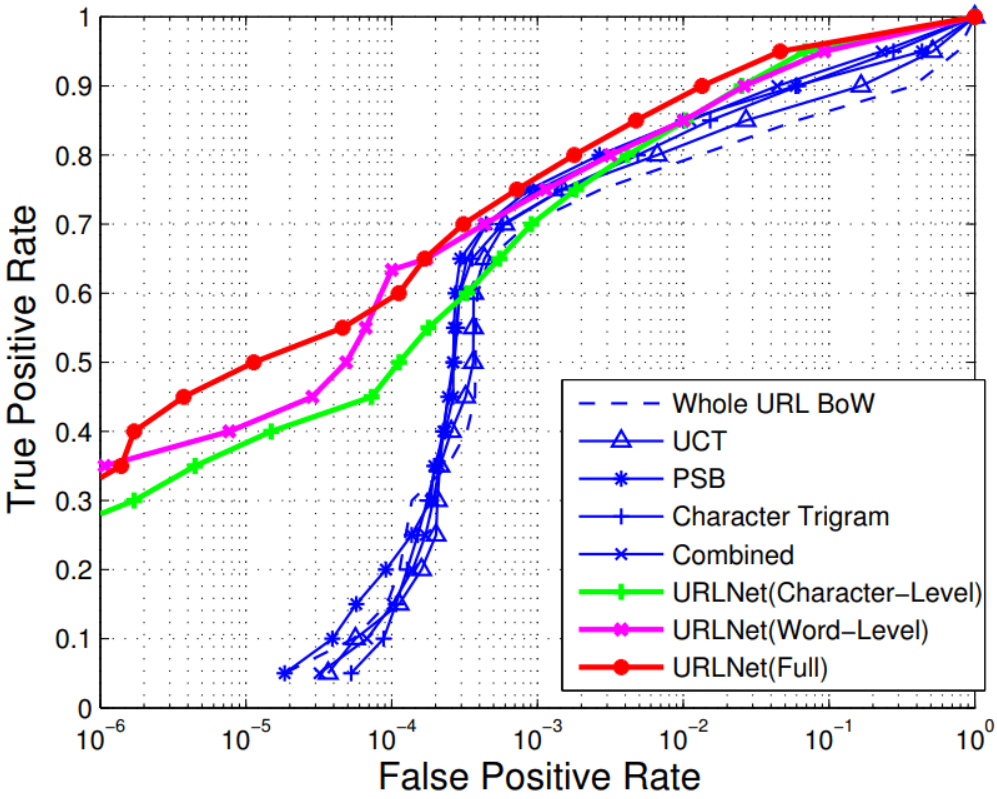

Рисунок 6 – ROC-кривая из статьи Лэ и других авторов за 2018 год

Учитывая, что ChatGPT в этих экспериментах работала в режиме zero-shot, то есть без обучения на примерах, результаты поразительны. Это использование ИИ эффективно для обнаружения фишинга, если модель хорошо натренировать. [28]

Выбор модели ИИ

Для эффективного обнаружения фишинговых атак используются различные модели искусственного интеллекта (ИИ), каждая из которых имеет свои преимущества в зависимости от задачи. Среди популярных моделей выделяются трансформеры, такие как BERT, и модели на основе градиентного бустинга, например, LightGBM.

Модель BERT (Bidirectional Encoder Representations from Transformers), разработанная Google, зарекомендовала себя как мощный инструмент для анализа текстов благодаря своей способности учитывать контекст слов в обе стороны. Модификации BERT, такие как RoBERTa и DistilBERT, обеспечивают еще более высокую точность. В задаче обнаружения фишинга такие модели могут эффективно распознавать сложные паттерны фишинговых сообщений и сайтов, но требуют значительных вычислительных ресурсов. Применение **BERT** в реальных системах, таких как у Яндекса для определения фишинговых сайтов, позволяет достигать точности до 99,9%.

С другой стороны, модели на основе градиентного бустинга, такие как LightGBM, тоже показывают хорошие результаты, особенно при работе с табличными данными, такими как URL или метаданные. Эти модели быстро обучаются и могут обрабатывать запросы за несколько миллисекунд, однако их точность ниже по сравнению с трансформерами, и они менее эффективны при работе с неструктурированными текстовыми данными.

Большие языковые модели, такие как GPT-**4**, также нашли свое применение в детектировании фишинговых атак. В исследованиях, таких как «ChatSpamDetector: Leveraging Large Language Models for Effective Phishing Email Detection», показано, что GPT-4 может достичь точности 99,7% при обнаружении фишинговых писем, но требует значительных вычислительных мощностей и более длительного времени обработки.

Таким образом, для создания системы обнаружения фишинга оптимальным выбором будет модель, которая сочетает высокую точность и достаточную скорость обработки. Модели на основе трансформеров, такие как BERT и RoBERTa, предпочтительны для задач, требующих глубокого понимания текста и контекста. В то время как LightGBM может быть хорошим выбором для быстрых решений с меньшими вычислительными затратами, но с компромиссами по точности.

BERT и трансформеры

Трансформеры – это архитектура нейросетей, применяемая в обработке естественного языка (NLP). Они работают на основе механизма *self-attention* (самовнимания), позволяющего анализировать контекст слова с учетом всех остальных слов в предложении. В отличие от рекуррентных сетей (RNN) трансформеры обрабатывают весь текст одновременно, что делает их более эффективными.

Основные компоненты трансформера:

- Self-attention – вычисляет вес слов относительно друг друга.
- Feed-forward слои – обрабатывают информацию после механизма внимания.
- Embeddings – представляют слова в числовом формате.
- Многоголовое внимание (multi-head attention) – анализирует разные контексты одновременно.
- Нормализация и пропускные соединения – помогают стабилизировать обучение.

BERT – это модель на основе трансформеров, разработанная Google. Ее ключевая особенность – двунаправленное обучение: модель анализирует текст как слева направо, так и справа налево, что улучшает понимание контекста.

Как работает BERT**:**

1. Входные данные: текст преобразуется в токены (единицы текста).
2. Embeddings: токены конвертируются в векторы.
3. Трансформер-энкодеры: несколько слоев обрабатывают информацию через механизмы самовнимания.
4. Выходные данные: полученные представления используются для различных NLP-задач (анализ тональности, машинный перевод, чат-боты).

BERT обучается с помощью двух задач:

- Masked Language Model (MLM) – часть слов в предложении заменяется на [MASK], и модель учится предсказывать их.

- Next Sentence Prediction (NSP) – модель определяет, связаны ли два предложения друг с другом.

На рисунке 7 изображен принцип работы BERT, где есть входные данные – два предложения, в которых по слову замаскировано, и BERT должен предсказать их. Также он должен определить, связаны ли между собой два предложения. Для работы слова, начало и разделения (точки) преобразуются в векторное представление токенов, а также пишется разделение предложений на A (первое) и B (второе). И также проставляется порядок слов. [33]

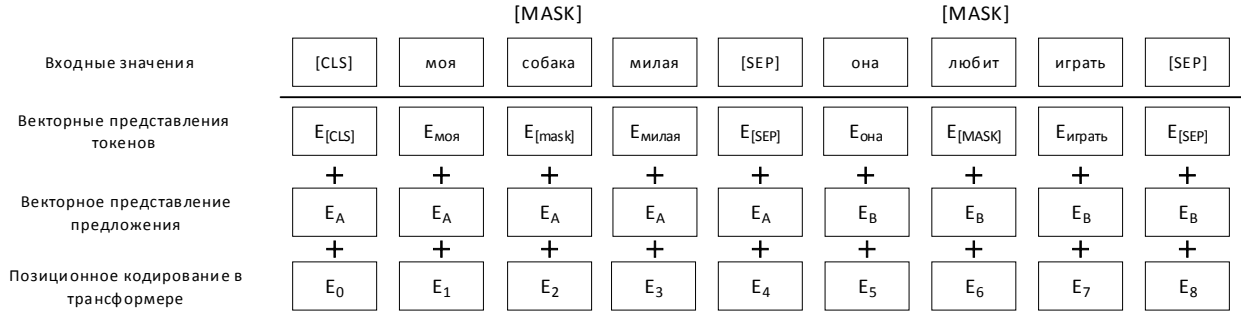

Рисунок 7 – Принцип работы BERT

Заключение

Подводя итоги проведенного аналитического обзора, можно говорить о том, что процесс разработки подсистемы SOC для защиты от фишинговых атак с применением технологий искусственного интеллекта и машинного обучения является актуальным и востребованным направлением в кибербезопасности. В ходе исследования были рассмотрены традиционные методы обнаружения фишинга, выявлены их ограничения и проведен анализ современных подходов на основе ИИ. Проведенное сравнение коммерческих решений показало, что готовые системы обладают высокой эффективностью, но требуют значительных финансовых вложений и адаптации под конкретные бизнес-процессы.

В ходе анализа моделей ИИ было установлено, что наиболее перспективным решением являются трансформерные архитектуры, в частности BERT и его модификации, которые демонстрируют высокую точность в выявлении фишинговых атак. Использование современных алгоритмов обработки естественного языка и методов машинного обучения позволяет адаптировать системы обнаружения фишинга к новым угрозам, повышая их эффективность в реальных условиях.

Следует отметить, что предложенный в работе подход к разработке собственной подсистемы SOC на основе ИИ представляется перспективным направлением, обеспечивающим эффективную защиту от фишинговых угроз, адаптивность к новым видам атак и снижение нагрузки на специалистов по кибербезопасности. Несмотря на необходимость первоначальных затрат на разработку и обучение моделей, в долгосрочной перспективе такие решения способны повысить уровень защиты информационных систем и оптимизировать процессы реагирования на инциденты.

Литература

1. Спам и фишинг в 2024 году – https://securelist.ru/spam-and-phishing-report-2024/111743/?utm_source=chatgpt.com (дата обращения 10.01.2025г.)

2. Число фишинговых сайтов, имитирующих финансовые бренды, выросло почти на 50% – https://cybersecurefox.com/ru/rost-fishinga-finansoviy-sektor-2024/ (дата обращения 12.01.2025г.)

3. Актуальные киберугрозы в странах СНГ 2023—2024 – https://www.ptsecurity.com/ru-ru/research/analytics/aktualnye-kiberugrozy-v-stranah-sng-2023-2024/?utm_source=chatgpt.com (дата обращения 15.01.2025г.)

4. ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ АЛГОРИТМОВ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ ЗАЩИТЫ ОТ ФИШИНГОВЫХ АТАК – https://cyberleninka.ru/article/n/issledovanie-vozmozhnostey-algoritmov-glubokogo-obucheniya-dlya-zaschity-ot-fishingovyh-atak?utm_source=chatgpt.com (дата обращения 20.01.2025г.).

5. PhishLabs – https://www.phishlabs.com/ (дата обращения 25.01.2025г.).

6. Cofense – https://cofense.com/ (дата обращения 27.01.2025г.).

7. Barracuda – https://www.barracuda.com/products/email-protection/plans (дата обращения 30.01.2025г.).

8. Proofpoint – https://www.proofpoint.com/us (дата обращения 10.02.2025г.).

9. KnowBe4 – https://www.knowbe4.com/ (дата обращения 12.02.2025г.).

10. Mimecast – https://www.mimecast.com/products/email-security/ (дата обращения 15.02.2025г.).

11. Barracude vs KnowBe4 – https://www.gartner.com/reviews/market/security-awareness-computer-based-training/compare/barracuda-vs-knowbe4?utm_source=chatgpt.com (дата обращения 20.02.2025г.).

12. Barracuda Security Awareness Training Alternatives – https://www.gartner.com/reviews/market/security-awareness-computer-based-training/vendor/barracuda/product/barracuda-security-awareness-training/alternatives (дата обращения 25.02.2025г.).

13. Attention Is All You Need - https://arxiv.org/abs/1706.03762 (дата обращения 28.02.2025г.).

14. RoBERTa: A Robustly Optimized BERT Pretraining Approach - https://arxiv.org/abs/1907.11692 (дата обращения 30.02.2025г.).

15. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter – https://arxiv.org/abs/1910.01108 (дата обращения 05.03.2025г.).

16. XLNet: Generalized Autoregressive Pretraining for Language Understanding – https://arxiv.org/abs/1906.08237 (дата обращения 07.03.2025г.).

17. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer – https://arxiv.org/abs/1910.10683 (дата обращения 10.03.2025г.).

18. GPT-4 - https://openai.com/index/gpt-4-research/ (дата обращения 12.03.2025г.).

19. Language Models are Few-Shot Learners - https://arxiv.org/abs/2005.14165 (дата обращения 15.03.2025г.).

20. Enriching Word Vectors with Subword Information – https://arxiv.org/abs/1607.04606 (дата обращения 18.03.2025г.).

21. ERNIE: Enhanced Representation through Knowledge Integration – https://arxiv.org/abs/1904.09223 (дата обращения 20.03.2025г.).

22. LightGBM: A Highly Efficient Gradient Boosting Decision Tree – https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf (дата обращения 20.03.2025г.).

23. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - https://arxiv.org/abs/1810.04805(дата обращения 21.03.2025г.).

24. Глубокое обучение в борьбе с фишингом: как ИИ распознает угрозы – https://cisoclub.ru/glubokoe-obuchenie-v-borbe-s-fishingom-kak-ii-raspoznaet-ugrozy/?utm_source=chatgpt.com (дата обращения 22.03.2025г.).

25. Исследование возможностей алгоритмов глубокого обучения для защиты от фишинговых атак - https://www.sberbank.ru/ru/person/kibrary/experts/issledovanie-vozmozhnostej-algoritmov-glubokogo-obucheniya-dlya-zashchity-ot-fishingovykh-atak (дата обращения 23.03.2025г.).

26. AI-Driven Phishing Detection Systems – https://www.researchgate.net/publication/382917933_AI-Driven_Phishing_Detection_Systems (дата обращения 24.03.2025г.).

27. Защита от фишинга с ИИ (искусственного интеллекта) – https://domznaniya.ru/page/zashchita-ot-fishinga-s-ii-iskusstvennogo-intellekta-4972291607/ (дата обращения 26.03.2025г.).

28. Что ChatGPT знает о фишинге? – https://securelist.ru/chatgpt-anti-phishing/107389/(дата обращения 05.03.2025г.). (дата обращения 26.03.2025г.).

29. Зоопарк трансформеров: большой обзор моделей от BERT до Alpaca – https://habr.com/ru/companies/just_ai/articles/733110/ (дата обращения 27.03.2025г.).

30. Роль ИИ в обеспечении ИБ: Яндекс.Толока и модель BERT для выявления фишинговых атак - https://baskobrin.ru/blog/rol-ii-v-obespechenii-ib-yandeks-toloka-i-model-bert-dlya-vyyavleniya-fishingovyh-atak (дата обращения 27.03.2025г.).

31. PhishNet: A Phishing Website Detection Tool using XGBoost – https://arxiv.org/abs/2407.04732 (дата обращения 28.03.2025г.).

32. ChatSpamDetector: Leveraging Large Language Models for Effective Phishing Email Detection – https://arxiv.org/abs/2402.18093(дата обращения 28.03.2025г.).

33. BERT в двух словах: Инновационная языковая модель для NLP – https://habr.com/ru/companies/otus/articles/702838/ (дата обращения 29.03.2025г.).