ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. №4 (32), Том 2. С. 204–211
DOI: https://doi.org/10.5281/zenodo.24132174
Сайт журнала: https://cajournal.kz

УДК 004.056:004.8

AI-СИСТЕМА ОБНАРУЖЕНИЯ СОЦИАЛЬНОЙ ИНЖЕНЕРИИ В ЦИФРОВЫХ КОММУНИКАЦИЯХ: МНОГОУРОВНЕВАЯ АРХИТЕКТУРА И МЕТОДИКА ЕЕ ЭКСПЕРИМЕНТАЛЬНОЙ ОЦЕНКИ

А.Р. Сарыкхазыева, А.Е. Магипарова
Satbayev University, Алматы, Казахстан

Сарыкхазыева Анеля Русланқызы — студентка 4 курса, кафедра «Кибербезопасность, обработка, и хранение информации», Satbayev University, Алматы, Казахстан.

Магипарова Айдана Ерболовна — студентка 4 курса, кафедра «Кибербезопасность, обработка, и хранение информации», Satbayev University, Алматы, Казахстан.

© А.Р. Сарыкхазыева, А.Е. Магипарова, 2026

Аннотация: В статье рассматривается задача автоматического выявления атак социальной инженерии (фишинг, целевой фишинг, компрометация деловой переписки) в электронной почте и мессенджерах. Показано, что распространенные детекторы, опирающиеся на стиль и лексику текста, теряют надежность при появлении сообщений, сгенерированных большими языковыми моделями, а также при переписке на нескольких языках. Предложена архитектура многоуровневой системы, объединяющей три независимых канала анализа: лингвистический анализ текста трансформерной моделью, оценку отклонения сообщения от привычного профиля общения отправителя и технический анализ адресов, заголовков и вложений. Описано правило слияния решений, корректно учитывающее отсутствие истории переписки. Разработана методика экспериментальной оценки на открытых корпусах писем, включающая кросс-корпусное тестирование и полный факторный эксперимент 2^3 для анализа влияния доли сообщений, сгенерированных языковой моделью, длины истории отправителя и веса лингвистического канала на качество обнаружения. Работа носит проектный характер: программный прототип и численные результаты испытаний являются предметом следующего этапа исследования.

Ключевые слова: социальная инженерия, фишинг, обнаружение атак, обработка естественного языка, трансформер, информационная безопасность, планирование эксперимента.

Введение

Социальная инженерия остается одним из самых результативных способов проникновения в организации, поскольку направлена не на уязвимости программного обеспечения, а на поведение человека. По данным отчета Verizon о расследовании утечек данных за 2024 год, человеческий фактор присутствовал примерно в двух третях проанализированных инцидентов [1]. Под социальной инженерией понимают совокупность приемов психологического воздействия, с помощью которых злоумышленник побуждает человека раскрыть конфиденциальные сведения, совершить платеж или выполнить иное действие в интересах атакующего [2]. Основой таких приемов служат универсальные механизмы влияния: ссылка на авторитет, создание дефицита времени, взаимность, социальное доказательство [3].

Масштаб проблемы подтверждается данными правоохранительных органов. По сведениям Центра жалоб на интернет-преступления ФБР (IC3), фишинг и подмена

отправителя в 2023 году оставались наиболее часто регистрируемой категорией жалоб, а потери от компрометации деловой переписки оценивались в несколько миллиардов долларов [4]. Ситуация осложняется тем, что большие языковые модели позволяют автоматически составлять убедительные целевые письма на основе открытых сведений о получателе, снижая стоимость атаки [5]. Тексты перестают содержать грамматические ошибки и шаблонные обороты, которые раньше служили надежным признаком обмана. Для Казахстана, где активно развиваются цифровые государственные и банковские сервисы, угроза усиливается особенностями языковой среды: деловая переписка ведется на русском, казахском и английском языках, нередко с переключением между ними внутри одного сообщения.

Автоматические системы защиты должны быть одновременно точными, устойчивыми к новым сценариям атак и понятными для человека, принимающего решение. Существующие решения, как показывают обзоры [6], обычно удовлетворяют лишь части этих требований: одни хорошо блокируют известные рассылки, но пропускают узконаправленные письма, другие анализируют только текст и не учитывают контекст общения.

Цель настоящей работы – обосновать архитектуру многоуровневой AI-системы обнаружения социальной инженерии в цифровых коммуникациях и разработать методику ее экспериментальной оценки. Для достижения цели решаются следующие задачи: анализ существующих подходов и выявление их ограничений; формулировка проверяемой гипотезы; проектирование модулей системы и правила объединения их решений; выбор корпусов данных, метрик и плана эксперимента.

1. Анализ существующих подходов к обнаружению социальной инженерии

Методы выявления вредоносных сообщений можно условно разделить на четыре группы в зависимости от того, какую информацию они используют.

Правила, списки и проверка подлинности отправителя. Классические фильтры используют черные списки адресов и доменов, сигнатуры и жесткие правила. К ним примыкают механизмы проверки отправителя SPF [7], DKIM [8] и DMARC [9], которые позволяют обнаружить подмену домена. Такие средства быстры и прозрачны, однако бессильны против письма, написанного вручную для конкретного сотрудника и отправленного с легитимного, но скомпрометированного или похожего домена.

Классическое машинное обучение. Ранние работы сравнивали логистическую регрессию, метод опорных векторов, случайный лес и другие алгоритмы на признаках, извлеченных из текста письма и его заголовков [10]. Отдельное направление анализирует лексические особенности URL без загрузки страницы, что снижает задержку и риск [11]. Недостатком является зависимость от ручного подбора признаков и быстрое устаревание модели при смене тактики атакующих.

Модели обработки естественного языка на основе трансформеров. Архитектура Transformer [12] и предобученные модели семейства BERT [13] учитывают контекст всего сообщения и улавливают не только отдельные слова-маркеры, но и структуру требования, скрытое давление на адресата. Многоязычные модели, например XLM-RoBERTa [14], применимы к текстам на нескольких языках. Общий недостаток этой группы – качество сильно зависит от обучающей выборки, а стилистические признаки теряют значимость, когда текст генерируется языковой моделью.

Контекстно-поведенческий анализ. Эта группа сравнивает сообщение с типичным профилем общения: частотой контактов, временем отправки, обычными темами и суммами, ролью отправителя. Она эффективна против компрометации деловой переписки, но требует доступа к истории коммуникаций и порождает вопросы приватности [6].

Сопоставление групп методов приведено в таблице 1.

© Central Asian Scientific Journal, 2026

Таблица 1 – Сравнение основных подходов к обнаружению социальной инженерии

Подход	Источник признаков	Достоинства	Ограничения
Правила и проверка отправителя [7–9]	Адреса, домены, заголовки	Простота, скорость, прозрачность	Не обнаруживает новые и целевые атаки
Классическое машинное обучение [10; 11]	Признаки текста, URL, заголовков	Умеренные требования к ресурсам, интерпретируемость	Ручной подбор признаков, устаревание
Трансформеры [12–14]	Семантика и контекст сообщения	Улавливают скрытое давление, работают с несколькими языками	Нужны данные и ресурсы, слабы против текстов языковых моделей
Контекстно-поведенческий анализ [6]	История переписки, профиль отправителя	Выявляет необычные запросы и подмену (BEC)	Требует истории, вопросы приватности

Анализ литературы позволяет выделить четыре нерешенные проблемы. Во-первых, детекторы, обученные на одном корпусе писем, заметно теряют точность на сообщениях другого происхождения; в обзорах [6] это называют проблемой обобщения. Во-вторых, лингвистические признаки обесцениваются, когда злоумышленник использует языковую модель для генерации текста [5]. В-третьих, большинство решений ориентировано на английский язык, а для русскоязычной и казахоязычной переписки открытых размеченных корпусов недостаточно. В-четвертых, пользователь обычно получает лишь итоговую метку без объяснения причин, что снижает доверие к результату и мешает обучению сотрудников.

2. Постановка задачи и гипотеза исследования

Объектом исследования являются цифровые сообщения (электронная почта и мессенджеры), содержащие текст, адреса, ссылки и вложения. Предметом исследования являются методы извлечения и объединения лингвистических, контекстно-поведенческих и технических признаков социальной инженерии.

Научная проблема состоит в противоречии между быстрым развитием средств генерации убедительного текста и недостаточной устойчивостью существующих детекторов, опирающихся преимущественно на стиль и содержание сообщения.

Рабочая гипотеза: объединение трех независимых групп признаков – лингвистических, контекстно-поведенческих и технических – позволит повысить качество обнаружения по сравнению с детектором, использующим только текст, причем выигрыш будет наибольшим на сообщениях, сгенерированных языковой моделью, где стилистические признаки теряют различающую способность, а несоответствия контекста и технические индикаторы сохраняются. Гипотеза является предположением и подлежит экспериментальной проверке по методике, изложенной в разделе 4.

3. Архитектура системы

Система проектируется как сервис, который получает копии входящих сообщений через почтовый шлюз или интеграцию с мессенджером, оценивает риск социальной инженерии и передает пользователю заключение с пояснением. Для определенности будем называть ее AntiSE (рабочее название). Обобщенная архитектура показана на рисунке 1.

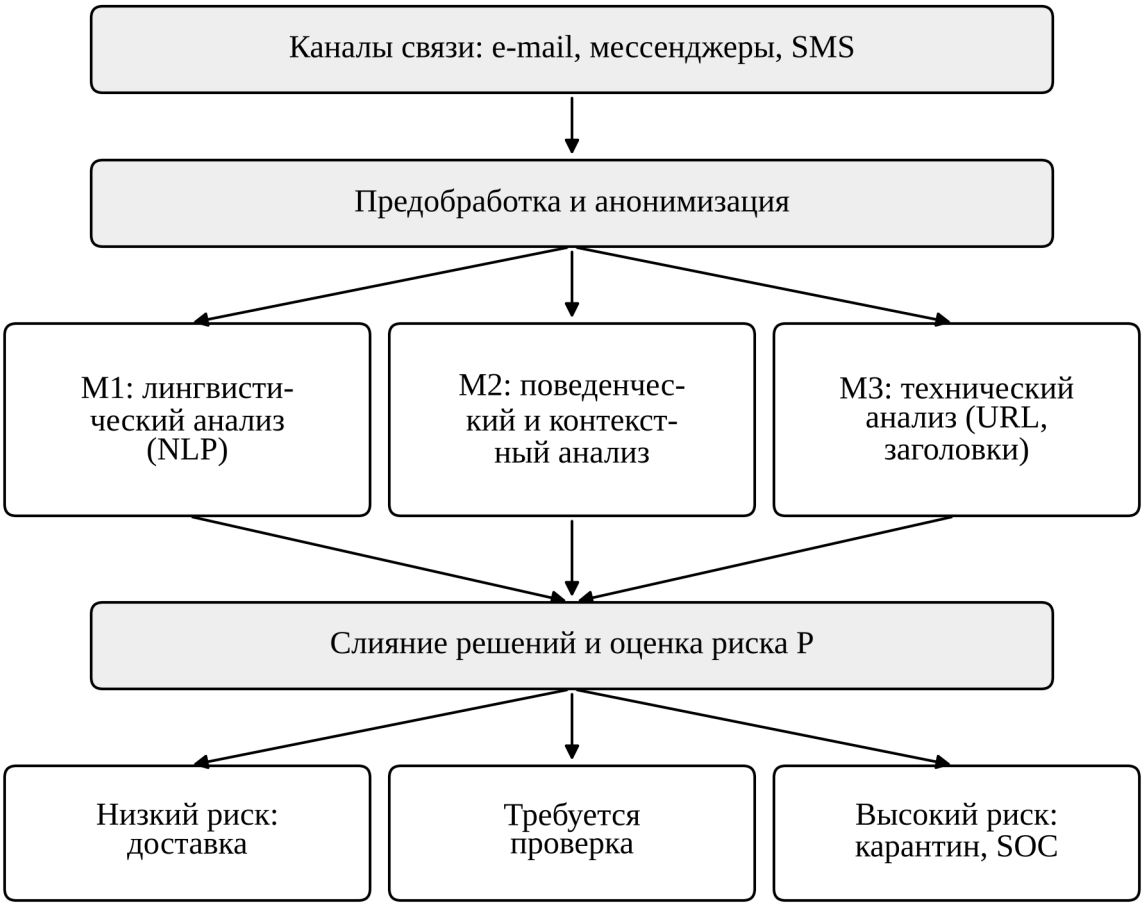


Рисунок 1 – Обобщенная архитектура системы AntiSE (разработано авторами)

Модуль предобработки декодирует сообщение, выделяет текст, ссылки, вложения и заголовки, нормализует текст и заменяет персональные данные (имена, номера счетов, телефоны) маркерами, чтобы дальнейший анализ не раскрывал лишних сведений. Язык сообщения определяется автоматически, допускается смешанный текст.

Модуль M1 (лингвистический анализ) использует многоязычный трансформерный кодировщик, например XLM-RoBERTa [14], дообученный на размеченных примерах. Модель оценивает вероятность p_l того, что сообщение содержит манипуляцию: требование срочности, угрозу, обещание выгоды, просьбу о конфиденциальности или об обходе принятых процедур.

Модуль M2 (контекстно-поведенческий анализ) сравнивает сообщение с профилем отправителя и получателя, накопленным по предыдущей переписке. Для каждого признака профиля (время отправки, длина сообщения, число получателей, наличие вложения, упоминание денежной суммы) вычисляется нормированное отклонение от среднего, а общая мера отклонения определяется выражением:

$$A = (1/K) \cdot \Sigma_k \min(1, |x_k - \mu_k| / (3\sigma_k)), \quad (1)$$

где K – число признаков, x_k – значение k-го признака в проверяемом сообщении, μ_k и σ_k – среднее и стандартное отклонение признака по истории отправителя. Дополнительно многомерная аномальность вектора признаков оценивается методом изолирующего леса [15], дающим оценку s. Итоговая вероятность модуля вычисляется как $p_c = (A + s) / 2$. Кроме того, проверяется расхождение адресов From и Reply-To, а также роль отправителя в структуре организации.

Модуль M3 (технический анализ) анализирует технические признаки: возраст и репутацию домена, результаты проверок SPF, DKIM и DMARC [7–9], различие между отображаемым и фактическим адресом ссылки, схожесть домена с доменами известных

организаций, тип и расширение вложения. Модель на основе градиентного бустинга [16] возвращает вероятность p_t.

Модуль слияния решений формирует итоговую оценку риска как взвешенную сумму оценок модулей:

$$P = (w_l \cdot p_l + m \cdot w_c \cdot p_c + w_t \cdot p_t) / (w_l + m \cdot w_c + w_t), \quad (2)$$

где w_l, w_c, w_t – веса модулей, подбираемые на валидационной выборке; m – индикатор достаточности истории (m = 1, если от отправителя получено не менее n_{min} сообщений, иначе m = 0). Нормировка в знаменателе важна: отсутствие истории не должно трактоваться ни как признак обмана, ни как признак легитимности – система опирается на оставшиеся каналы и сообщает пользователю об ограничении.

Итоговая оценка P сопоставляется с двумя порогами $\tau_1 < \tau_2$ и переводится в один из трех уровней: «низкий риск» (сообщение доставляется), «требуется проверка» (сообщение сопровождается предупреждением) и «высокий риск» (сообщение помещается в карантин и передается аналитику центра мониторинга). Пользователь получает не только уровень риска, но и вклад каждого модуля, а для лингвистического канала – слова и фразы, повлиявшие на решение, выделенные методом SHAP [17]. Такое заключение носит вероятностный характер и не может рассматриваться как самостоятельное доказательство вредоносности сообщения. Решения аналитиков и жалобы пользователей возвращаются в обучающую выборку. Последовательность обработки сведена в таблицу 2.

Таблица 2 – Этапы обработки сообщения в системе AntiSE

Этап	Входные данные	Результат
1. Прием сообщения	Письмо или сообщение мессенджера	Сообщение и метаданные
2. Предобработка	Текст, заголовки, ссылки, вложения	Нормализованный текст, признаки, индикатор m
3. Модуль M1	Нормализованный текст	Оценка p_l, выделенные фразы
4. Модуль M2	Профиль отправителя и получателя	Оценка p_c
5. Модуль M3	Домены, заголовки, вложения	Оценка p_t
6. Слияние	p_l, p_c, p_t, m	Итоговая оценка P и уровень риска
7. Отчет	P и данные модулей	Заключение для пользователя или аналитика

4. Методика экспериментальной оценки

Данные. Для оценки планируется использовать открытый корпус деловой переписки Enron [18] в качестве источника легитимных сообщений и публичный корпус фишинговых писем Nazario [19]; для обучения модуля M3 дополнительно привлекаются проверенные фишинговые адреса из базы PhishTank [20]. Для проверки на русском и казахском языках предполагается сформировать собственный размеченный корпус из обезличенных и синтетических примеров. Часть фишинговых писем для тестовой выборки генерируется открытой языковой моделью по сценариям, заданным авторами, в изолированной среде; такие письма используются только для оценки детектора [5]. Разбиение на обучающую, валидационную и тестовую выборки выполняется по времени и по отправителям: одно и то же лицо не должно встречаться в разных частях выборки, иначе оценка качества окажется завышенной.

Метрики. Качество классификации оценивается точностью (Precision), полнотой (Recall), F1-мерой и долей ложных срабатываний (FPR):

$$\text{Precision} = TP / (TP + FP), \quad \text{Recall} = TP / (TP + FN), \quad (3)$$

$$F1 = 2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall}), \quad FPR = FP / (FP + TN), \quad (4)$$

где TP – верно обнаруженные вредоносные сообщения, FP – легитимные сообщения, ошибочно признанные вредоносными, FN – пропущенные вредоносные сообщения, TN –

верно пропущенные легитимные. Для сервиса защиты важны оба вида ошибок: пропуск целевого письма может привести к прямому ущербу, а избыток ложных тревог снижает доверие и приводит к игнорированию предупреждений. Дополнительно фиксируются задержка обработки одного сообщения и потребление памяти.

Сравнение вариантов. Эксперимент проводится по схеме последовательного наращивания: базовый лингвистический детектор (M1), затем M1 + M3 и полная система M1 + M2 + M3. Все варианты проверяются на одной и той же тестовой выборке. Для оценки обобщающей способности выполняется кросс-корпусный тест: обучение на одном корпусе и проверка на другом без дообучения.

Факторный эксперимент. Чтобы проверить ключевую часть гипотезы – что контекстные и технические каналы сохраняют полезность на сообщениях, сгенерированных языковой моделью, – планируется полный факторный эксперимент 2^3 [21]. Откликом служит F1-мера, факторами – доля сгенерированных писем среди вредоносных в тестовой выборке, минимальная длина истории отправителя и вес лингвистического канала. Уровни факторов приведены в таблице 3.

Таблица 3 – Факторы и уровни варьирования

Фактор	Нижний уровень (−1)	Верхний уровень (+1)
x_1 – доля сообщений, сгенерированных языковой моделью, %	10	60
x_2 – минимальная длина истории n_{min}, сообщений	5	30
x_3 – вес лингвистического канала w_l	0,2	0,6

План эксперимента включает 8 опытов – все сочетания уровней трех факторов, каждый опыт повторяется трижды на разных подвыборках тестовых данных. По результатам строится уравнение регрессии с учетом взаимодействий:

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3, \quad (5)$$

где коэффициенты вычисляются как $b_i = (1/8)\cdot\Sigma\ x_{iu}\cdot\bar{y}_u$, а $\bar{y}_u$ – средний отклик в u-м опыте. Однородность дисперсий проверяется по критерию Кохрена, значимость коэффициентов – по критерию Стьюдента, адекватность модели – по критерию Фишера. Ортогональность плана гарантирует, что оценки коэффициентов не зависят друг от друга.

Особый интерес представляет коэффициент взаимодействия b_{13}. Если гипотеза верна, он окажется значимым и отрицательным: увеличение веса лингвистического канала будет снижать F1 тем сильнее, чем больше доля сгенерированных сообщений. Значимый отрицательный b_1 при этом подтвердит деградацию качества на текстах языковых моделей, а знак b_2 покажет, окупается ли требование более длинной истории потерей охвата новых отправителей.

5. Обсуждение и ограничения

Предлагаемый подход опирается на идею, что разные виды признаков теряют информативность в разной степени при смене тактики атакующего. Стиль текста легко имитировать, тогда как расхождение с привычным поведением отправителя, несоответствие доменов и результаты проверок подлинности являются свойствами контекста, а не формулировок. Поэтому объединение каналов потенциально повышает устойчивость системы, а раздельные оценки модулей делают заключение более понятным для пользователя.

Вместе с тем у подхода есть ограничения. Модуль M2 работает только при наличии достаточной истории; для новых отправителей система опирается на два оставшихся канала. Анализ содержимого сообщений затрагивает персональные данные, поэтому обработку целесообразно выполнять внутри периметра организации, с минимизацией хранимых сведений и журналированием доступа в соответствии с законодательством. Для казахского языка размеченных корпусов существенно меньше, чем для английского, что может снизить качество модуля M1 и потребует формирования собственных данных. Злоумышленники

могут адаптировать атаки под известный детектор, поэтому модели нуждаются в регулярном дообучении. Наконец, выделенные методом SHAP фразы показывают, на что опиралась модель, но не доказывают причинно-следственную связь с фактом атаки.

Важно подчеркнуть, что настоящая работа представляет проектное обоснование системы. Утверждать о превосходстве предлагаемого подхода над существующими решениями можно будет только после реализации прототипа и проведения описанных экспериментов. Автоматическое обнаружение следует рассматривать как дополнение, а не замену программ повышения осведомленности сотрудников.

Заключение

В работе проанализированы основные подходы к обнаружению социальной инженерии и показано, что их ключевыми ограничениями являются слабая обобщающая способность, потеря значимости стилистических признаков на текстах языковых моделей и недостаточная поддержка многоязычной переписки. Предложена архитектура многоуровневой системы AntiSE, объединяющей лингвистический, контекстно-поведенческий и технический анализ, и правило слияния решений, корректно учитывающее отсутствие истории отправителя. Разработана методика экспериментальной оценки, включающая кросс-корпусное тестирование и полный факторный эксперимент 2^3, позволяющий количественно оценить влияние доли сгенерированных сообщений, длины истории и веса лингвистического канала на качество обнаружения.

Дальнейшая работа включает формирование размеченного корпуса на русском и казахском языках, программную реализацию прототипа, обучение модулей, проведение запланированных экспериментов и проверку сформулированной гипотезы. Практическая ценность результатов состоит в создании инструмента, помогающего организациям и пользователям выявлять атаки социальной инженерии на ранней стадии.

Список литературы:

1. Verizon. 2024 Data Breach Investigations Report [Электронный ресурс]. – Verizon Business, 2024. – URL: https://www.verizon.com/business/resources/reports/dbir/ (дата обращения: 11.11.2026).

2. Hadnagy C. Social Engineering: The Science of Human Hacking. – 2nd ed. – Indianapolis: Wiley, 2018. – 320 p.

3. Cialdini R. B. Influence: The Psychology of Persuasion. – New York: Harper Business, 2006. – 336 p.

4. Federal Bureau of Investigation. Internet Crime Report 2023 [Электронный ресурс]. – Internet Crime Complaint Center (IC3), 2024. – URL: https://www.ic3.gov/ (дата обращения: 11.11.2026).

5. Hazell J. Spear Phishing With Large Language Models [Электронный ресурс] // arXiv. – 2023. – arXiv:2305.06972. – URL: https://arxiv.org/abs/2305.06972 (дата обращения: 11.11.2026).

6. Salloum S., Gaber T., Vadera S., Shaalan K. A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques // IEEE Access. – 2022. – Vol. 10. – P. 65703–65727.

7. Kitterman S. Sender Policy Framework (SPF) for Authorizing Use of Domains in Email, Version 1: RFC 7208 [Электронный ресурс]. – IETF, 2014. – URL: https://www.rfc-editor.org/rfc/rfc7208 (дата обращения: 11.11.2026).

8. Crocker D., Hansen T., Kucherawy M. DomainKeys Identified Mail (DKIM) Signatures: RFC 6376 [Электронный ресурс]. – IETF, 2011. – URL: https://www.rfc-editor.org/rfc/rfc6376 (дата обращения: 11.11.2026).

9. Kucherawy M., Zwicky E. Domain-based Message Authentication, Reporting, and Conformance (DMARC): RFC 7489 [Электронный ресурс]. – IETF, 2015. – URL: https://www.rfc-editor.org/rfc/rfc7489 (дата обращения: 11.11.2026).

10. Abu-Nimeh S., Nappa D., Wang X., Nair S. A Comparison of Machine Learning Techniques for Phishing Detection // Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit. – 2007. – P. 60–69.

11. Sahingoz O. K., Buber E., Demir O., Diri B. Machine Learning Based Phishing Detection from URLs // Expert Systems with Applications. – 2019. – Vol. 117. – P. 345–357.

12. Vaswani A., Shazeer N., Parmar N. [et al.]. Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.

13. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT 2019. – 2019. – P. 4171–4186.

14. Conneau A., Khandelwal K., Goyal N. [et al.]. Unsupervised Cross-lingual Representation Learning at Scale // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). – 2020. – P. 8440–8451.

15. Liu F. T., Ting K. M., Zhou Z.-H. Isolation Forest // Proceedings of the 8th IEEE International Conference on Data Mining (ICDM). – 2008. – P. 413–422.

16. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. – P. 785–794.

17. Lundberg S. M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 4765–4774.

18. Klimt B., Yang Y. The Enron Corpus: A New Dataset for Email Classification Research // Machine Learning: ECML 2004. – Berlin: Springer, 2004. – (Lecture Notes in Computer Science; Vol. 3201). – P. 217–226.

19. Nazario J. Phishing Corpus [Электронный ресурс]. – URL: https://monkey.org/~jose/phishing/ (дата обращения: 11.11.2026).

20. PhishTank. Developer Information: Verified Phishing URLs [Электронный ресурс]. – URL: https://phishtank.org/ (дата обращения: 11.11.2026).

21. Montgomery D. C. Design and Analysis of Experiments. – 9th ed. – Hoboken: Wiley, 2017. – 640 p.