ТЕХНИКАЛЫҚ ҒЫЛЫМДАР
ТЕХНИЧЕСКИЕ НАУКИ
TECHNICAL SCIENCE

CENTRAL ASIAN SCIENTIFIC JOURNAL
ISSN 3135-3061
2026. №4 (32), Том 2. С. 276–281
DOI: https://doi.org/10.5281/zenodo.24186299
Сайт журнала: https://cajournal.kz

УДК 004.056.5:004.8

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ОБНАРУЖЕНИЯ ФИШИНГОВЫХ ЭЛЕКТРОННЫХ ПИСЕМ

Д.К. Набоков, Р.В. Ичетовкин, Н.А. Ломако, А.Ю. Баширов
Казахский национальный исследовательский технический университет имени К. И. Сатпаева, Алматы, Казахстан

Набоков Даниэль Кириллович — студент, кафедра кибербезопасности, Казахский национальный исследовательский технический университет имени К. И. Сатпаева, Алматы, Казахстан.

Ичетовкин Роман Валерьевич — студент, кафедра кибербезопасности, Казахский национальный исследовательский технический университет имени К. И. Сатпаева, Алматы, Казахстан.

Ломако Никита Александрович — студент, кафедра кибербезопасности, Казахский национальный исследовательский технический университет имени К. И. Сатпаева, Алматы, Казахстан.

Баширов Артур Юрьевич — студент, кафедра кибербезопасности, Казахский национальный исследовательский технический университет имени К. И. Сатпаева, Алматы, Казахстан.

© Д.К. Набоков, Р.В. Ичетовкин, Н.А. Ломако, А.Ю. Баширов, 2026

Аннотация: В работе рассматривается проектирование интеллектуальной системы распознавания фишинговых электронных писем на основе совместного анализа их содержания и технических характеристик. Цель исследования состоит в обосновании архитектуры, которая объединяет семантические модели обработки естественного языка, результаты аутентификации электронной почты и признаки URL-адресов. На основании анализа научных публикаций и стандартов SPF, DKIM и DMARC предложены схема обработки сообщения, состав признаков, порядок формирования оценки риска и протокол проверки качества. Особое внимание уделено происхождению технических данных, снижению числа пропущенных фишинговых сообщений, многоязычной переписке и интерпретируемости решения. Сопоставлены достоинства и ограничения правил, классического машинного обучения и трансформерных моделей. Результатом является концептуальная спецификация гибридного детектора, которую можно использовать при последующей программной реализации. Работа носит проектно-аналитический характер: обучение готовой модели и измерение её точности на реальном корпусе электронных писем не проводились.

Ключевые слова: фишинговые письма, информационная безопасность, машинное обучение, трансформерные модели, анализ URL, SPF, DKIM, DMARC, оценка риска.

Введение

Электронная почта используется для деловой переписки, передачи документов, подтверждения финансовых операций и восстановления доступа к цифровым сервисам. Вместе с этим она остаётся одним из каналов социальной инженерии: злоумышленник стремится заставить получателя перейти по поддельной ссылке, раскрыть пароль или выполнить действие от имени доверенного контрагента. Фишинговое письмо может имитировать официальный стиль организации, использовать адрес отправителя, визуально похожий на настоящий, и содержать ссылку на сайт с похожим доменным именем.

Традиционная фильтрация писем опирается на сигнатуры, блокирующие списки, репутацию домена и эвристические правила. Подобный подход остаётся полезным для массовых известных кампаний, однако испытывает затруднения при появлении новых шаблонов, персонализированных текстов и ранее не встречавшихся доменов. Обзор P. H. Kyaw и соавторов [1] связывает развитие направления с использованием глубокого обучения; сравнительная работа R. Meléndez и соавторов [2] показывает потенциал трансформерных моделей на конкретном наборе данных, но одновременно обнаруживает ошибки классификации сложных текстов.

Основная исследовательская проблема заключается в том, что текст, техническая подлинность отправителя и характеристики ссылок обладают разной природой. Письмо может пройти проверку аутентификации домена и при этом содержать обманное предложение; напротив, легитимное сообщение иногда не проходит отдельные проверки из-за особенностей пересылки или настройки почтовой инфраструктуры. Поэтому ни один отдельный показатель не следует рассматривать как достаточное доказательство фишинга либо его отсутствия. Цель статьи — разработать обоснованную схему гибридного анализа, задающую последовательность обработки письма, состав признаков и критерии последующей экспериментальной проверки.

Сравнительный анализ методов

Для обнаружения фишинга применяют три группы методов. Правила и сигнатуры легко объяснить оператору и быстро выполнить, но требуют регулярного обновления. Классические алгоритмы машинного обучения — логистическая регрессия, опорные векторы, случайный лес — выявляют статистические связи в текстовых и инженерных признаках. Их результат существенно зависит от набора данных, качества разметки и выбранного способа представления текста. Трансформерные языковые модели учитывают контекст слов и способны распознавать более сложные сценарии социально-инженерного воздействия [2, 3].

Применение трансформеров связано с дополнительными затратами на обучение и выполнение предсказаний. Кроме того, высокий результат на одном тестовом корпусе не гарантирует устойчивости к новым фишинговым кампаниям. Исследование C. Opara и соавторов [4] обращает внимание на особенности фишинговых писем, сформированных с применением генеративных моделей. Для подобных писем орфографические и стилистические ошибки уже не являются надёжным сигналом, поэтому целесообразно дополнять анализ текста проверкой инфраструктурных признаков.

Таблица 1 — Сопоставление подходов к обнаружению фишинговых писем

Подход	Сильная сторона	Ограничение
Правила и списки	Быстрый и прозрачный контроль известных шаблонов	Слабая адаптация к новым атакам
TF-IDF и классические ML-модели	Невысокие вычислительные затраты, воспроизводимость	Зависимость от словаря и обучающей выборки
BERT/RoBERTa-подобные модели	Контекстный анализ смысла письма	Ресурсоёмкость и риск ошибок на новых доменах
Гибридный анализ	Сопоставление независимых текстовых и технических сигналов	Необходимость калибровки и проверки источника признаков

Таким образом, перспективна не замена всех проверок одной нейросетью, а согласованная обработка нескольких модальностей. Качество такого объединения определяется не только выбранным классификатором, но и тем, какие поля сообщения действительно получены доверенным почтовым сервером.

Архитектура предлагаемой системы

Входом системы служит электронное сообщение в формате MIME. На первом этапе выполняется безопасный разбор заголовков, текстовой и HTML-частей, вложений и гиперссылок. Обработка не должна автоматически выполнять сценарии, загружать удалённые изображения или переходить по адресам. HTML преобразуется в текстовое

представление с сохранением ссылок и признаков подозрительной разметки. После нормализации данные направляются в три логически независимых модуля.

Первый модуль выполняет семантический анализ темы и тела письма. В качестве базового варианта можно использовать TF-IDF и логистическую регрессию, а для исследования контекстной составляющей — дообученный DistilBERT либо другую трансформерную архитектуру. Модели BERT и DistilBERT подробно описаны в [5, 6]. Целевой выход модуля — оценка вероятности фишингового сценария и признаки, позволившие сформировать эту оценку: запрос кода подтверждения, ложная срочность, имитация службы поддержки или попытка изменить реквизиты оплаты.

Второй модуль исследует технические данные. В качестве признаков используются согласованность полей From, Reply-To и Return-Path, результаты SPF, DKIM и DMARC и сведения о том, где именно получены результаты проверок. Стандарты SPF [7], DKIM [8] и DMARC [9] задают взаимодополняющие механизмы проверки отправителя. При этом текстовое поле Authentication-Results, вставленное недоверенным узлом, нельзя автоматически считать достоверным. Источник такого поля и границы доверия к почтовому шлюзу должны быть определены отдельно [10].

Третий модуль анализирует URL и домены без автоматического открытия вредоносных страниц. В расчёт могут приниматься количество и структура ссылок, наличие кодировки международных доменных имён, сходство с известными корпоративными доменами, расхождение видимого текста и фактического адреса, необычные цепочки перенаправлений и доступные репутационные сведения. Применение машинного обучения для выявления фишинговых URL исследовано, в частности, в работе O. K. Sahingoz и соавторов [11].

Результаты модулей объединяются в общий вектор признаков вместе с отметками их доступности и достоверности. Если, например, проверка DMARC технически недоступна, система должна фиксировать отсутствие результата, а не автоматически приписывать сообщению успешную или неуспешную проверку. Такой принцип уменьшает риск ошибочного решения при неполноте почтовых метаданных.



Рисунок 1 — Схема обработки входящего сообщения (разработка авторов)

В качестве принципа объединения может использоваться метаклассификатор, получающий текстовую оценку, технические и URL-признаки, а также индикаторы их доступности. Его выходом становится оценка риска p, лежащая в диапазоне от 0 до 1. Операционный сценарий предусматривает два настраиваемых порога: ниже первого письмо допускается к обычной доставке, между порогами направляется на дополнительную проверку, выше второго помещается в карантин. Конкретные пороги определяются только после оценки ошибок на отдельной валидационной выборке.

Для оператора полезно выводить краткое объяснение: «домен ссылки отличается от домена отправителя», «обнаружен запрос одноразового кода», «не пройдена проверка DMARC на доверенном шлюзе». Объяснимость не означает, что любое локальное

объяснение достоверно отражает весь процесс классификации. Методы интерпретации, в том числе SHAP, следует применять с проверкой их устойчивости и качества [12].

План экспериментальной проверки

Предлагаемая архитектура является проектной спецификацией, поэтому следующие положения относятся к плану исследований, а не к уже полученным экспериментальным результатам. Для объективной оценки потребуется корпус легитимных и фишинговых сообщений с документированными источниками и датами. Дубликаты, почти одинаковые тексты, повторяющиеся шаблоны и письма одной кампании должны группироваться до разделения данных, иначе тест может завышать результат из-за утечки содержания между обучением и проверкой.

Рекомендуется сравнить не менее четырёх конфигураций: базовую TF-IDF с логистической регрессией; текстовый трансформер; текстовую модель с техническими признаками; полную гибридную модель с модулем URL. Для каждого варианта должны быть одинаковы условия подготовки тестовых данных, а параметры моделей следует выбирать только на обучающей и валидационной частях. Отдельно требуется проверить письма на русском и казахском языках, а также смешанные русско-английские сообщения. Многоязычное представление текста может строиться на современных кросс-языковых моделях, исследованных в [13].

В качестве ключевых показателей следует использовать полноту обнаружения фишинга (Recall), точность положительных решений (Precision), F1-меру, долю ложных пропусков (FNR), площадь под Precision–Recall-кривой и время обработки одного письма. Особенно важно оценивать систему при реальной доле фишинга в потоке, поскольку обычная Accuracy может маскировать ошибки на редком опасном классе. Методические основания для использования Precision–Recall-графиков на несбалансированных данных приведены в [14].

Таблица 2 — Проверяемые гипотезы и критерии оценки предлагаемой системы

Проверяемая гипотеза	Сравнение	Критерий оценки
H1. Технические признаки дополняют текст	Текстовая модель / текст + аутентификация	Recall, F1, число ложных пропусков
H2. URL-признаки повышают устойчивость	Модель без URL / полная модель	PR-AUC на внешних доменах
H3. Учет доступности признаков снижает ошибки	Без индикаторов / с индикаторами	Калибровка и ошибки при пропусках данных
H4. Многоязычная проверка выявляет слабые места	Обычный тест / русский и казахский тест	Recall по языковым подгруппам

Для использования в почтовом шлюзе необходимо отдельно измерить среднюю задержку и 95-й процентиль времени обработки, число сообщений, требующих ручной проверки, и устойчивость к отсутствию внешнего репутационного сервиса. Важна калибровка вероятностей, то есть согласованность выдаваемых рисков с наблюдаемыми частотами ошибок. Понятие и оценку калибровки нейросетевых классификаторов рассматривают C. Guo и соавторы [15]. Без такой проверки значение p не следует интерпретировать как подтверждённую статистическую вероятность атаки.

Практическое применение и ограничения

Для организаций Казахстана существенна поддержка русскоязычной, казахскоязычной и смешанной переписки. Технические признаки позволяют частично компенсировать недостаток размеченных текстов отдельных языков, однако не устраняют проблему полностью. Промышленный детектор также должен обрабатывать пересылаемые письма, рассылки через сторонние сервисы, сообщения от скомпрометированных реальных аккаунтов и временную недоступность систем репутационной проверки. При внедрении необходимо соблюдать требования к защите персональных данных и минимизации хранения содержимого писем.

Отдельного внимания требует обратная связь от аналитиков. Исправленные ими метки пригодны для улучшения модели только после контроля качества и исключения повторов из тестового набора. Автоматическое дообучение на непрерывном потоке без контроля создаёт риск накопления ошибок и внесения злоумышленником отравленных примеров. Следовательно, полезно разделять эксплуатационный мониторинг, периодическую оценку качества и процедуру выпуска новой версии модели.

Ограничения настоящей работы принципиальны: предложенная система пока не реализована в виде завершённого программного комплекса; проверка на новом размеченном корпусе не проведена; точность, полнота и время обработки не измерены. Числовые значения из отдельных исследований [2–4] относятся исключительно к соответствующим авторам и наборам данных. Представленная схема служит технически обоснованной основой для следующего этапа — разработки прототипа и независимой валидации.

Заключение

В статье предложена структура интеллектуальной системы обнаружения фишинговых электронных писем на основе объединения семантического анализа, проверки технических заголовков и исследования URL-адресов. Центральная идея заключается в раздельном получении признаков, учёте их происхождения и доступности, формировании объяснимой оценки риска и маршрутизации сомнительных сообщений на дополнительную проверку. Сопоставление методов показывает, что контекстные языковые модели и технические проверки выполняют разные функции, поэтому их целесообразно рассматривать как взаимодополняющие компоненты. Сформулирован протокол будущего исследования, предусматривающий проверку на внешних и многоязычных данных, контроль утечек, анализ ошибок и измерение задержки. Достоверность предлагаемых улучшений должна быть подтверждена последующими экспериментами.

Список литературы:

1. Kyaw P. H., Gutierrez J., Ghobakhlou A. A Systematic Review of Deep Learning Techniques for Phishing Email Detection // Electronics. — 2024. — Vol. 13, № 19. — Art. 3823. — DOI: 10.3390/electronics13193823.

2. Meléndez R., Ptaszynski M., Masui F. Comparative Investigation of Traditional Machine-Learning Models and Transformer Models for Phishing Email Detection // Electronics. — 2024. — Vol. 13, № 24. — Art. 4877. — DOI: 10.3390/electronics13244877.

3. Alhuzali A., Alloqmani A., Aljabri M., Alharbi F. In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning and Deep Learning Models Across Multiple Datasets // Applied Sciences. — 2025. — Vol. 15, № 6. — Art. 3396. — DOI: 10.3390/app15063396.

4. Opara C., Modesti P., Golightly L. Evaluating Spam Filters and Stylometric Detection of AI-Generated Phishing Emails // Expert Systems with Applications. — 2025. — Vol. 276. — Art. 127044. — DOI: 10.1016/j.eswa.2025.127044.

5. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. — 2019. — P. 4171–4186. — DOI: 10.18653/v1/N19-1423.

6. Sanh V., Debut L., Chaumond J., Wolf T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter // arXiv. — 2019. — URL: https://arxiv.org/abs/1910.01108 (дата обращения: 13.11.2026).

7. Kitterman S. Sender Policy Framework (SPF) for Authorizing Use of Domains in Email, Version 1. RFC 7208. IETF, 2014. — DOI: 10.17487/RFC7208.

8. Crocker D., Hansen T., Kucherawy M. DomainKeys Identified Mail (DKIM) Signatures. RFC 6376. IETF, 2011. — DOI: 10.17487/RFC6376.

9. Herr T., Levine J. Domain-Based Message Authentication, Reporting, and Conformance (DMARC). RFC 9989. IETF, 2026. — URL: https://www.rfc-editor.org/rfc/rfc9989.html (дата обращения: 13.11.2026).

10. Kucherawy M. Message Header Field for Indicating Message Authentication Status. RFC 8601. IETF, 2019. — DOI: 10.17487/RFC8601.

© Central Asian Scientific Journal, 2026

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. — DOI: 10.1016/j.eswa.2018.09.029.

12. Lundberg S. M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems. — 2017. — Vol. 30. — URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (дата обращения: 13.11.2026).

13. Conneau A. et al. Unsupervised Cross-lingual Representation Learning at Scale // Proceedings of ACL. — 2020. — P. 8440–8451. — DOI: 10.18653/v1/2020.acl-main.747.

14. Saito T., Rehmsmeier M. The Precision–Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets // PLOS ONE. — 2015. — Vol. 10, № 3. — Art. e0118432. — DOI: 10.1371/journal.pone.0118432.

15. Guo C., Pleiss G., Sun Y., Weinberger K. Q. On Calibration of Modern Neural Networks // Proceedings of ICML. — 2017. — P. 1321–1330. — URL: https://proceedings.mlr.press/v70/guo17a.html (дата обращения: 13.11.2026).

© Central Asian Scientific Journal, 2026