Модели машинного обучения для выявления целевых атак через вложения в электронных письмах

К.А. Беляков, А.В. Шевченко, Е.М. Гиндин

МИРЭА - Российский технологический университет

Аннотация: В статье рассматривается задача выявления вредоносных вложений в электронных письмах, используемых при целевых кибератаках. Предложен подход, основанный на совместном использовании текстовых и файловых признаков сообщений с применением методов машинного обучения. Выполнено сравнение моделей логистической регрессии и метода случайного леса по основным метрикам качества классификации. Эксперименты на синтетическом наборе данных показали, что логистическая регрессия обеспечивает более высокую полноту обнаружения вредоносных вложений, тогда как случайный лес характеризуется более высокой точностью классификации. Полученные результаты подтверждают эффективность мультипризнаковой модели и возможности интеграции в системы защиты электронной почты.

Ключевые слова: машинное обучение, целевая атака, электронная почта, фишинг, вредоносное вложение, обнаружение атак, информационная безопасность.

Введение. В последние годы наблюдается тенденция к усложнению и повышению целенаправленности кибератак, что требует применения усовершенствованных подходов к их раннему выявлению. Электронная почта остаётся основным вектором распространения вредоносных вложений, в том числе при реализации целевой фишинговой атаки и атак типа деловой компрометации электронной почты [1]. Классические методы защиты, основанные на сигнатурном анализе, не обеспечивают должной эффективности при обнаружении ранее неизвестных угроз, что обуславливает необходимость применения интеллектуальных методов анализа данных [2].

Целью исследования является разработка и экспериментальная оценка модели машинного обучения, обеспечивающей обнаружение вредоносных вложений в электронных письмах на основе совокупности текстовых и файловых признаков.

Для достижения поставленной цели решались следующие задачи:

1. Провести анализ существующих подходов к выявлению вредоносных вложений;

2. Сформировать набор данных, моделирующий структуру реальных электронных сообщений;

3. Разработать и обучить модели машинного обучения;

4. Провести сравнительный анализ эффективности алгоритмов классификации;

5. Определить наиболее информативные признаки для идентификации атакующих писем.

Для дальнейшего обоснования выбора методологии исследования необходимо провести анализ существующих подходов к обнаружению вредоносных вложений и оценить их применимость при выявлении целевых атак. Рассмотрим основные подходы, используемые для обнаружения вредоносных вложений в электронных письмах, и проанализируем их эффективность.

Сигнатурный анализ является одним из наиболее распространённых и традиционных методов защиты от вредоносных вложений, лежащим в основе работы большинства антивирусных решений [3]. Данный подход основан на сравнении исследуемого файла с базой известных сигнатур, характеризующих конкретные образцы вредоносного программного обеспечения. Основным преимуществом сигнатурного метода является высокая точность и скорость при обнаружении уже известных угроз. Однако его существенным недостатком является неэффективность при выявлении ранее не встречавшихся (нулевого дня) образцов, а также полиморфных и метаморфных вирусов, способных изменять свой код для обхода сигнатурных баз [4]. Это ограничивает применимость данного подхода при обнаружении целевых атак, которые, как правило, используют новые или модифицированные вредоносные объекты.

В ряде исследований отмечается, что первые антиспам-системы начала 2000-х годов базировались преимущественно на сигнатурном анализе, что обеспечивало высокую точность при обнаружении уже известных угроз, однако не позволяло своевременно выявлять новые атаки [5].

Для повышения адаптивности и точности при обнаружении фишинговых сообщений и вредоносных вложений применяются алгоритмы машинного обучения классического типа, включая наивный байесовский классификатор, метод опорных векторов (SVM) и случайный лес [5]. Указанные методы зарекомендовали себя как эффективные инструменты для решения задач классификации текстовых и структурированных данных, обеспечивая возможность комплексного анализа статистических и содержательных признаков электронных писем. Эффективность таких моделей в значительной степени определяется качеством подготовки обучающей выборки, корректностью извлечения признаков и параметризацией алгоритмов. Вместе с тем, при появлении новых типов атак и изменении распределения данных наблюдается снижение точности классификации, обусловленное чувствительностью моделей к смещению выборки и дисбалансу классов, что ограничивает их применение в условиях изменчивого профиля кибератак [6].

Развитие методов искусственного интеллекта привело к широкому внедрению глубоких нейронных сетей. Для обработки содержимого электронных писем и вложений применяются сверточные нейронные сети (convolutional neural networks - CNN), рекуррентные архитектуры (recurrent neural networks - RNN), а также современные модели, такая как двунаправленная модель представления текста на основе трансформеров (Bidirectional Encoder Representations Transformers - BERT) и его модификации [7]. Таким образом, эти подходы обеспечивают высокую способность к обобщению, автоматическое выделение значимых признаков и

устойчивость к большей вариативности данных. В ряде исследований отмечается значительное превосходство трансформерных моделей над традиционными методами классификации при решении задач антифишинга.

Несмотря на значительный прогресс, достигнутый в области противодействия фишинговым и целевым атакам, существующие подходы характеризуются рядом методологических и практических ограничений [8]. Сигнатурные методы обнаружения не обладают способностью к идентификации ранее неизвестных или модифицированных вредоносных объектов, что снижает их эффективность в условиях изменяющихся тактик атакующих. Алгоритмы классического машинного обучения демонстрируют зависимость от качества извлечения признаков и репрезентативности обучающих данных, а применение моделей глубокого обучения сопряжено с высокими вычислительными издержками и необходимостью наличия обширных размеченных выборок. Перечисленные факторы определяют актуальность разработки гибридных и интеграционных подходов, сочетающих преимущества различных классов методов и обеспечивающих более точное выявление целевых атак, реализуемых посредством вредоносных вложений в электронных письмах.

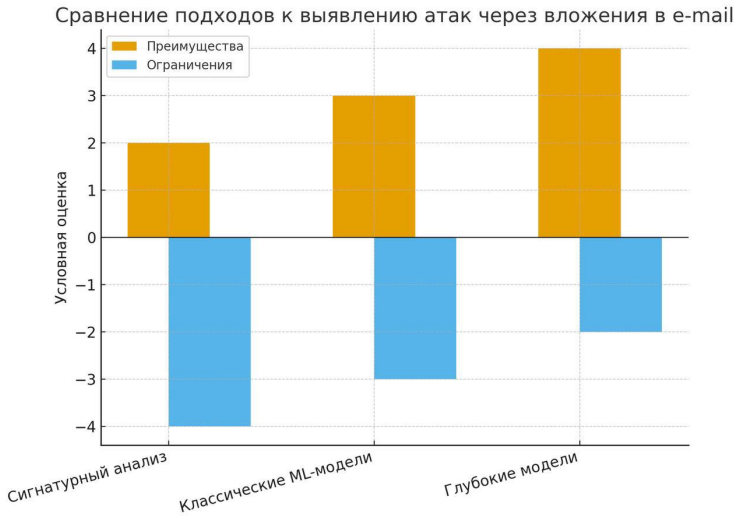

Рис. 1. – Классификация методов обнаружения вредоносных вложений

Таблица № 1

Сравнительная таблица методов обнаружения фишинга

№	Подход	Преимущества	Ограничения
1	Сигнатурный анализ	Высокая точность при обнаружении известных угроз (Четкость>0.98). Минимальные вычислительные затраты (реализация на уровне потокового анализа). Простая интеграция в антивирусные и почтовые фильтры.	Не выявляет атаки нулевого дня и модифицированные образцы. Требует постоянного обновления сигнатурных баз. Невозможность обобщения - алгоритм работает только с совпадениями байт/хэшей.
2	Классические ML-модели (случайный лес, логистическая регрессия)	Учитывают статистические и семантические признаки (энтропия, наличие макросов). Поддерживают интерпретацию результатов и важности признаков. Высокая скорость	Зависимость от качества извлечения признаков и балансировки классов. Ограниченная способность к генерализации при изменении

		обучения и предсказания (O(n·log n) или меньше). Возможность адаптации под конкретный домен данных.	распределения данных. Уязвимость к состязательным атакам (изменение текста или структуры вложения снижает точность)
3	Глубокие нейронные сети	Автоматическое выделение признаков без ручного проектирования. Высокая обобщающая способность, возможность анализа семантики текста писем и структуры вложений. Поддержка мультимодального анализа (текст + бинарные признаки). Стабильное качество при наличии большого объёма данных (F1-	Большие требования к вычислительным ресурсам (GPU, ≥ 8 GB VRAM). Необходимость больших размеченных датасетов (> 10 000 образцов). Сложность интерпретации решений (проблема черного ящика). Потенциальные риски переобучения при ограниченном наборе данных.

		мера>0.9).	
4	Гибридные методы	Объединяют преимущества разных классов алгоритмов — устойчивость к новым типам атак и к дисбалансу классов. Возможность объединения текстовых и бинарных признаков (multi-feature fusion). Повышенная устойчивость к фальсифицированным данным за счёт ансамблирования.	Более высокая вычислительная сложность ($O(n^2)$ при ансамблях из нескольких моделей). Необходимость настройки гиперпараметров и валидации на множестве конфигураций. Требует экспертных знаний при интеграции в почтовые шлюзы и SIEM-системы.

Методика исследования заключается в организации и проведении экспериментов, направленных на проверку эффективности предложенного подхода. Для проведения экспериментов был сформирован синтетический набор данных, моделирующий реальные электронные письма. Каждое письмо включало тему, тело, вложение и бинарную метку класса («вредоносное» или «нормальное»).

При генерации учитывались:

- вероятностное распределение расширений вложений (docx, pdf, zip, exe и др.);

- ключевые слова, характерные для атакующих писем (срочно, пароль, перейти по ссылке, выигрыш, конфиденциально - urgent, password, click, win, confidential);
- структурные признаки вложений: размер файла, энтропия содержимого, наличие макросов [9].

Около 20% сообщений формировались как вредоносные, что отражает реальную дисбалансированную природу задач кибербезопасности.

Перед построением моделей машинного обучения текстовые и файловые данные были приведены к единому формату с помощью комплекса предобработки, обеспечивающего корректное извлечение признаков и совместный анализ. Для анализа текста письма (тема и тело) применялось представление взвешенных терминов TF-IDF (term frequency - inverse document frequency) с ограничением словаря до 400 признаков.

Расширения вложений кодировались с помощью бинарного кодирования категориальных признаков (one-hot encoding), а также сохранялись числовые характеристики: размер вложения (в КБ), энтропия содержимого, бинарный признак «наличие макроса». В результате, итоговый вектор признаков включал как лексические, так и структурные характеристики.

Для оценки качества классификации электронных писем были использованы две модели машинного обучения: Случайный лес (Random Forest - RF) и Логистическая регрессия (Logistic Regression - LR).

В таблице 1 представлены основные метрики: точность (Accuracy), точность предсказания положительного класса, полнота (Recall), среднее гармоническое значение F1-мера и ROC-кривая.

Для количественной оценки эффективности построенных моделей использовался набор стандартных метрик классификации, позволяющих комплексно характеризовать их точность и способность к выявлению

целевых атак: точность (Accuracy), четкость (Precision), полнота (Recall), F1-мера, ROC-кривая. Формулы представлены ниже:

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

$$Recall = \frac{TP}{TP + FN}, \quad F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}.$$

Рис. 2. – Формулы качества моделей

Таблица № 2

Сравнение качества моделей

Модель	Общая точность	Точность	Полнота	F1-мера	площадь под ROC-кривой
Случайный лес	0.818	0.846	0.11	0.195	0.843
Логистическая регрессия	0.788	0.482	0.81	0.604	0.868

Анализ показывает, что:

Случайный лес достигает высокой общей точности (0.818) и четкость (0.846), что говорит о низком числе ложных срабатываний. Однако показатель полноты (0.11) крайне низкий, что означает, что модель плохо выявляет вредоносные письма.

Логистическая регрессия демонстрирует более сбалансированные результаты: полноты достигает 0.81, что значительно лучше выявляет вредоносные вложения, хотя точность ниже (0.482).

Для более наглядной оценки были построены матрицы ошибок для обеих моделей (рисунок 3).

Низкое значение полноты при высокой ROC-кривой объясняется особенностями выбранного порога классификации, при котором модель демонстрирует хорошую ранжирующую способность, но смещается в сторону минимизации ложных срабатываний [1].

Из рисунка видно, что случайный лес в основном классифицирует письма как «нормальные», пропуская вредоносные. Логистическая регрессия лучше улавливает вредоносные сообщения, но допускает больше ошибок при классификации нормальных.

Рис. 3. – Матрица ошибок для случайного леса и логистической регрессии.

По графику видно, что обе модели демонстрируют качество выше случайного угадывания (линия диагонали). Логистическая регрессия имеет более высокую площадь под ROC-кривой = 0.868, что подтверждает её преимущество в данной задаче.

Для выявления наиболее информативных факторов, влияющих на классификацию сообщений, был проведён анализ значимости признаков на основе модели случайного леса. В таблице 3 представлены 15 наиболее значимых слов, влияющих на классификацию писем.

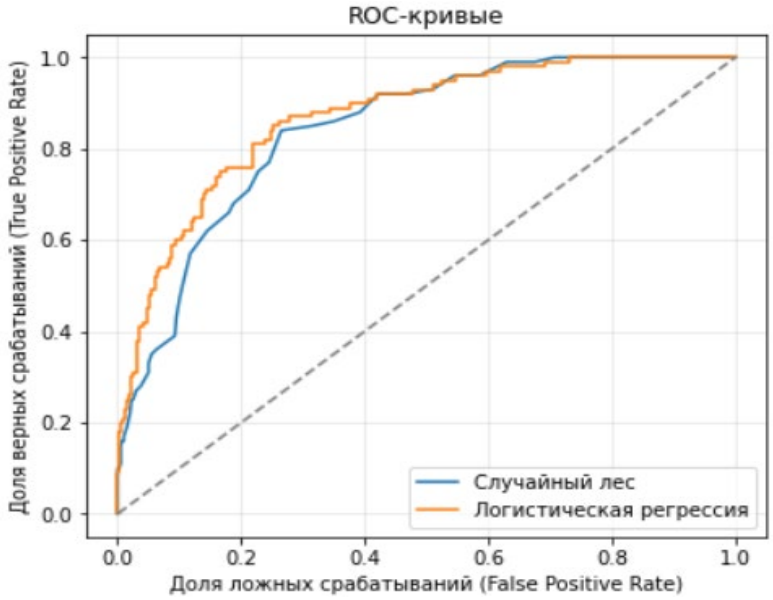

Рис. 4. – ROC-кривые для моделей

Таблица № 3

Топ-15 признаков по важности (Случайный лес)

Признак	Важность
urgent	0.0547
execute	0.0546
regards	0.0524
schedule	0.0522
link	0.0473
project	0.0473
confidential	0.0468
payment	0.0459
win	0.0457
attached	0.0455
document	0.0443
password	0.0428
thanks	0.0414
report	0.0402
team	0.0397

Исходя из данных таблицы 3, можно сделать вывод, что слова, связанные с финансовыми транзакциями (payment, confidential, win), документами (attached, document, password) и срочностью (urgent, execute), являются ключевыми индикаторами потенциально вредоносных сообщений.

Обсуждение результатов

Анализ экспериментальных данных показал, что модель логистической регрессии обеспечивает более высокую полноту классификации = 0.81 при умеренном снижении четкости = 0.48, что делает её предпочтительным инструментом для задач выявления вредоносных вложений, где критично минимизировать количество пропущенных угроз. В отличие от неё, модель случайного леса демонстрирует высокую общую точность классификации, однако характеризуется низкой чувствительностью к вредоносным объектам, что ограничивает её применимость в условиях, требующих оперативного обнаружения целевых атак. Полученные результаты указывают на необходимость выбора модели с учётом конкретных требований к балансировке полноты и точности, а также на потенциал комбинированных подходов, объединяющих преимущества различных алгоритмов для повышения надёжности и устойчивости систем защиты электронной почты.

Результаты коррелируют с данными зарубежных исследований, где отмечается преимущество линейных и нейросетевых моделей при анализе комбинированных признаков сообщений [7]. Использование синтетического датасета позволило контролировать распределение классов, однако дальнейшие исследования требуют привлечения реальных почтовых данных и учёта разнообразия форматов вложений.

Перспективным направлением дальнейшей работы является интеграция предложенной модели в инфраструктуру корпоративной информационной безопасности, включая системы мониторинга событий безопасности и почтовые шлюзы. Данное внедрение позволит реализовать раннее выявление целевых атак на этапе доставки сообщений, обеспечивая минимизацию вероятности компрометации пользовательских устройств. Кроме того, применение модели в составе автоматизированных систем обработки событий создаёт возможности для непрерывного обновления и адаптации

модели к новым и модифицированным типам угроз, что существенно повышает устойчивость защиты и эффективность проактивного реагирования на инциденты.

Заключение

В работе предложен подход к выявлению вредоносных вложений в электронной почте с использованием методов машинного обучения. Интеграция текстовых и файловых признаков позволила построить модель на основе комбинированных признаков, обеспечивающую повышение точности обнаружения атак по метрике F1-мера.

Научная новизна заключается в комплексном анализе сообщений, включающем как семантические характеристики текста, так и параметры вложений. Практическая значимость заключается в возможности применения результатов при создании интеллектуальных систем защиты почтовых серверов и корпоративных сетей.

Дальнейшие исследования планируется направить на расширение набора данных, внедрение нейросетевых архитектур на основе трансформеров и реализацию автоматического выявления вложений в реальном времени, также, в качестве одного из направлений расширения обучающей выборки может рассматриваться использование открытых датасетов вредоносного программного обеспечения, предназначенных для обучения и валидации моделей машинного обучения [10].

Литература (References)

1. Rudd E. M., Harang R., Saxe J., et al. MEADE: Towards a Malicious Email Attachment Detection Engine. — 2018. — ArXiv: 1804.08162. — URL: arxiv.org/abs/1804.08162 (дата обращения: 20.09.2025).

2. Maiorca D., Biggio B., Giacinto G. On the robustness of PDF malware detectors // USENIX Workshop on Offensive Technologies (WOOT). — 2019. —

URL: usenix.org/conference/woot19/presentation/maiorca (дата обращения: 22.09.2025).

3. Tzermias Z., Sykiotis C., Papadogiannakis A., Markatos E. P. Combining static and dynamic analysis for the detection of malicious documents // Proceedings of the 4th European Workshop on System Security (EUROSEC). — 2011. — URL: usenix.org/legacy/event/eurosec11/tech/final_files/Tzermias.pdf (дата обращения: 20.09.2025).

4. Maiorca D., Biggio B., Giacinto G. Towards Adversarial Malware Detection: Lessons Learned from PDF-based Attacks. — 2018. — ArXiv: 1811.00830. — URL: arxiv.org/abs/1811.00830 (дата обращения: 25.09.2025).

5. Nguyen T. T., Armitage G. A survey of techniques for Internet traffic classification using machine learning // IEEE Communications Surveys & Tutorials. — 2008. — T. 10, № 4. — pp. 56–76. — URL: ieeexplore.ieee.org/document/4624260 (дата обращения: 19.09.2025).

6. Anderson B., Paul S., McGrew D. Deciphering malware's use of TLS (without decryption). — 2016. — URL: arxiv.org/abs/1607.01639 (дата обращения: 17.09.2025).

7. Fettaya R., Mansour A. Detecting malicious PDF using Convolutional Neural Networks. — 2020. — ArXiv: 2007.12729. — URL: arxiv.org/abs/2007.12729 (дата обращения: 22.09.2025).

8. Mohammed T. M., Nataraj L., Chikkagoudar S., Chandrasekaran S., Manjunath B. S. HAPSSA: Holistic Approach to PDF Malware Detection Using Signal and Statistical Analysis. — 2021. — ArXiv: 2111.04703. — URL: arxiv.org/abs/2111.04703 (дата обращения: 22.09.2025).

9. Saha A., Lindorfer M. Exploring the malicious document threat landscape. — 2024. — URL: martina.lindorfer.in/files/papers/maldocs_worma24.pdf (дата обращения: 24.09.2025).

10. Anderson H. S., Roth P. EMBER: An open dataset for training static PE malware machine learning models. — 2018. — ArXiv: 1804.04637. — URL: arxiv.org/abs/1804.04637 (дата обращения: 25.09.2025).

Авторы согласны на обработку и хранение персональных данных.

Дата поступления: 27.11.2025

Дата публикации: 6.02.2026