DOI 10.37882/2223-2966.2025.04-2.05

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ФИЛЬТРАЦИИ СПАМ-КОНТЕНТА

USING MACHINE LEARNING TO FILTER SPAM CONTENT

N. Verezubova
N. Sakovich
A. Chekulaev

Summary. The article examines the effectiveness of the Random Forest machine learning method for classifying and filtering spam content. The study demonstrates the process of training a model on a labeled text corpus, the features of data preprocessing, and the extraction of significant features. The study includes training and measuring the model, as well as interpreting the results using metrics, including those reflecting accuracy on unbalanced data.

Keywords: random forest, machine learning, spam content, information security.

Верезубова Наталья Афанасьевна
Кандидат экономических наук, доцент,
Московская государственная академия ветеринарной медицины и биотехнологии имени К.И. Скрябина
nverez@mail.ru
Сакович Наталия Евгениевна
Доктор технических наук, доцент,
Брянский государственный аграрный университет
nasa2610@mail.ru
Чекулаев Артур Анатольевич
Московская государственная академия ветеринарной медицины и биотехнологии имени К.И. Скрябина

Аннотация. В статье рассматривается эффективность применения метода машинного обучения «случайный лес» (Random Forest) для классификации и фильтрации спам-контента. Исследование демонстрирует процесс обучения модели на размеченном корпусе текстов, особенности предобработки данных и выделения значимых признаков. В рамках исследования проводится обучение и измерение модели, а также трактовка результатов по метрикам, в том числе, отражающих точность на несбалансированных данных.

Ключевые слова: случайный лес, машинное обучение, спам контент, информационная безопасность.

Введение

Защита пользователя от вредоносного контента, включая спам-письма с потенциально опасным содержимым, представляет собой одну из фундаментальных целей в рамках комплексного кластера задач информационной безопасности. Такие критические аспекты данного направления как доступность, целостность и конфиденциальность в своей синергетической совокупности формируют безопасную экосистему для конечного пользователя. Данная триада безопасности не просто теоретический конструкт, но практический фундамент противодействия современным угрозам, включая вредоносные спам-рассылки [1].

Для эффективного создания и поддержания такой защищённой среды требуется имплементация многоуровневого подхода с применением широкого спектра инструментов. В современном контексте цифровой трансформации особую значимость приобретают передовые методики машинного обучения, демонстрирующие исключительную эффективность в обнаружении и нейтрализации вредоносных спам-писем и других угроз нового поколения. Алгоритмы глубокого обучения обладают уникальной способностью к выявлению сложных паттернов в потоках данных, что делает их незаменимыми в идентификации спам-сообщений с замаскированным вредоносным кодом и других полиморфных угроз [2].

Однако интеграция методов машинного обучения в системы информационной безопасности требует тщательного баланса между эффективностью защиты и сохранением приватности пользовательских данных. Федеративное обучение и дифференциальная приватность становятся ключевыми технологиями, позволяющими анализировать потенциально опасные спам-сообщения без необходимости централизованного хранения чувствительной информации пользователей [3].

В контексте постоянно эволюционирующего ландшафта киберугроз, где спам-письма с вредоносными вложениями и ссылками представляют серьёзную опасность, адаптивные системы защиты, основанные на принципах машинного обучения, представляют собой не просто технологическое преимущество, но стратегическую необходимость для обеспечения устойчивой безопасности цифровой инфраструктуры [4, 5].

Таким образом актуальность исследования обоснована необходимостью популяризации знаний о вну-

тренней составляющей «чёрного ящика» систем информационной безопасности, способных эффективно противодействовать современным формам спама и связанным с ним вредоносным программам.

Материалы и методы

Основным методом данного исследования выступит метод компьютерного моделирования, на основе публично доступного набора данных, содержащего как образцы электронных спам-писем (spam), так и не спам контента (ham).

Затем на основании данного датасета был обучен классификатор случайного леса (RF).

Классификатор случайного леса — это ансамблевый метод классификации данных, основанный на совмещении метода бэггинга и методе случайных деревьев [6].

Выбор алгоритма был обоснован высокой точностью классификатора в решении подобных задач [7].

Работа классификатора была оптимизирована при помощи решётчатого метода гиперпараметрического поиска, задававшего такие параметры случайного леса как критерий (gini, entropy), глубину дерева, максимальное количество образцов в листе, количество самих деревьев и создание bootstrap-выборки [8, 9].

Следует отметить, что датасет содержал крайне несбалансированные данные, разделение spam-ham показано на рисунке 1.

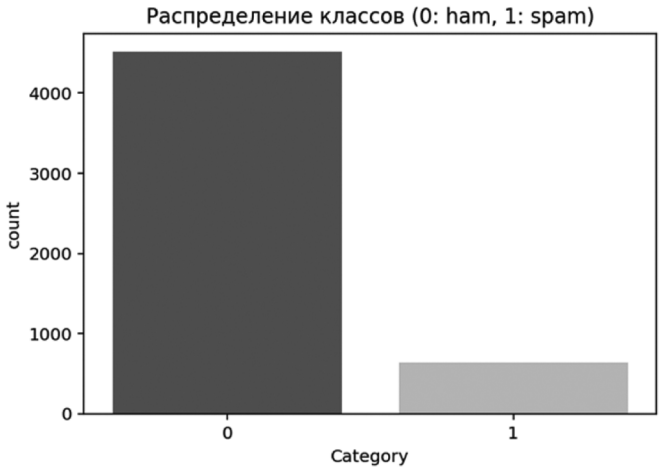

Рис. 1. Распределение spam-ham в обучающем датасете

В качестве борьбы с данным дисбалансом, и для качественной оценки работы модели был применён комплекс методов, таких как K-Fold Cross-Validation (на 5 фолдах) и взвешенность классов. Затем с модели были сняты такие метрики как F1-score, G-Mean (среднее геометрическое) и ROC-AUC.

Результаты и обсуждение

После обучения модели и осуществления её качественного измерения, результаты вычисления были занесены в результирующую таблицу, представленную в таблице 1.

Таблица 1.

Результаты работы модели

Метрика	Результат
G-Mean	0.9329
ROC-AUC	0.9800
F1-score	0.9427

Результаты работы алгоритма, исходя из полученных во время измерения данных можно сделать следующие выводы:

Во-первых: Такой показатель как среднее геометрическое, в данном случае используемое для оценки бинарной классификации, демонстрирует среднюю для модели случайного леса точность.

Во-вторых: Относительная площадь ROC-AUC указывает на высокую способность модели разделять предоставленные ей классы.

В-третьих: Показатель F1 метрики указывает на высокое сродство исходных и валидационных данных.

В общем можно сделать вывод, что модель, смоделированная с учётом дисбаланса в данных, качественно смогла разделить полученные данные и классифицировать их должным образом, что указывает на важность применения адекватных ситуации метрик и способов обучения.

Выводы

В рамках данного исследования была смоделирована система для классификации спам контента на основе классификатора случайного леса. Было измерено качество модели соответствующим несбалансированной обучающей выборке метрикам.

Результаты исследования показали, насколько важно правильно выбрать и методы исследования, и критерии оценки его результатов. Удачный выбор этих параметров позволил создать эффективную систему распознавания спама, подтвердив тем самым высокую эффективность алгоритма «случайного леса» для решения подобных задач.

В целом данное исследование подтверждает необходимость адекватного подбора сочетания методов эксперимента и его метрик. В свою очередь правильно подобранные условия привели к качественной работе классификатора, что подтверждает эффективность случайного леса в задачах классификации спам контента.

ЛИТЕРАТУРА

1. Артюхин, В.В. Информационная беззащитность / В. В. Артюхин // Прикладная информатика. — 2008. — № 6(18). — С. 32–43. — EDN IVYOHD.
2. Резниченко, Л.С. Методы машинного обучения в задаче обеспечения информационной безопасности / Л.С. Резниченко // Информационные ресурсы и системы в экономике, науке и образовании: сборник статей XIV Международной научно-практической конференции, Пенза, 29–30 апреля 2024 года. — Пенза: Автономная некоммерческая научно-методическая организация «Приволжский Дом знаний», 2024. — С. 143–148. — EDN BCVLWW.
3. Матыюк, С.П. Построение глубокой нейронной сети для задачи обнаружения кибератак / С.П. Матыюк // Актуальные проблемы науки и образования в условиях современных вызовов (шифр — МКАП 25) : Сборник материалов XXV Международной научно-практической конференции, Москва, 17 ноября 2023 года. — Москва: Печатный цех, 2023. — С. 52–60. — EDN WNRXHM.
4. Ивкина, М.С. Решение задачи классификации на основе случайного леса / М.С. Ивкина // Современные технологии в науке и образовании — СТНО-2018: Сборник трудов международного научно-технического форума: в 11 томах, Рязань, 28 февраля 2018 года / Под общ. ред. О.В. Миловзорова. Том 3. — Рязань: Рязанский государственный радиотехнический университет, 2018. — С. 61–65. — EDN XVSNFR.
5. Spam Email Classification using Random Forest // kaggle.com [Электронный ресурс]. — Режим доступа — URL: https://www.kaggle.com/code/ardava/spam-email-classification-using-random-forest (дата обращения: 04.04.2025).
6. Федорова С.А. Разработка спам-фильтра с использованием методов машинного обучения // Время науки — The Times of Science. 2023. №4-1. [Электронный ресурс]. — Режим доступа — URL: https://cyberleninka.ru/article/n/razrabotka-spam-filtra-s-ispolzovaniem-metodov-mashinnogo-obucheniya (дата обращения: 10.04.2025).
7. Самигулин Т.Р., Джурабаев А.Э.У. Анализ тональности текста методами машинного обучения // Научный результат. Информационные технологии. 2021. №1. С. 55–62.
8. Казаков М.А. Алгоритм кластеризации на основе разбиения пространства признаков // Вест. КРАУНЦ. Физ.-мат. науки. 2022. №2. [Электронный ресурс]. — Режим доступа — URL: https://cyberleninka.ru/article/n/algoritm-klasterizatsii-na-osnove-razbieniya-prostranstva-priznakov (дата обращения: 01.04.2025).
9. Михайлов И.С., Зеар Аунг, Йе Тху Аунг. Разработка модификации метода опорных векторов для решения задачи классификации с ограничениями на предметную область // Программные продукты и системы. 2020. №3. [Электронный ресурс]. — Режим доступа — URL: https://cyberleninka.ru/article/n/razrabotka-modifikatsii-metoda-opornyh-vektorov-dlya-resheniya-zadachi-klassifikatsii-s-ogranicheniyami-na-predmetnuyu-oblast (дата обращения: 01.04.2025).