УДК 004

Горюнова Д.А.

студент

Кузьмина Е.С.

студент

Поволжский государственный университет телекоммуникаций и информатики

ОБНАРУЖЕНИЕ СПАМА В ЭЛЕКТРОННОЙ ПОЧТЕ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ: СРАВНЕНИЕ ФУНКЦИОНАЛЬНОЙ РАЗРАБОТКИ И КОМПЛЕКСНЫХ МЕТОДОВ.

Аннотация: Работа посвящена анализу и сравнению различных подходов к обнаружению спама в электронной почте с использованием машинного обучения. В статье рассматриваются два основных типа методов: функциональная разработка и комплексные методы. Был проведен сравнительный анализ эффективности и точности различных алгоритмов машинного обучения. Результаты исследования демонстрируют преимущества и недостатки каждого подхода, а также выявляют наиболее эффективные методы для борьбы с современными формами спама. Статья предлагает рекомендации по выбору оптимального метода обнаружения спама в зависимости от конкретных условий и задач.

Ключевые слова: машинное обучение, сквозные методы глубокого обучения, традиционные методы обучения, обработка данных.

Goryunova D.A.

student

Kuzmina E.S.

student

Povolzhskiy ***State University of Telecommunications and Informatics***

Machine learning-based spam detection in e-mail: a comparison of functional development and complex methods.

Abstract: The paper is devoted to the analysis and comparison of various approaches to detecting spam in e-mail using machine learning. The article discusses two main types of methods: functional development and complex methods. A comparative analysis of the effectiveness and accuracy of various machine learning algorithms was carried out. The results of the study demonstrate the advantages and disadvantages of each approach, as well as identify the most effective methods to combat modern forms of spam. The article offers recommendations on choosing the optimal spam detection method, depending on specific conditions and tasks.

Keywords: machine learning, end-to-end deep learning methods, traditional learning methods, data processing.

Введение

Рассылаемые в больших количествах нежелательные сообщения, или электронный спам, составляют значительную часть мирового почтового трафика. Они приводят к потере времени, снижению производительности и возможным угрозам безопасности. Чтобы уменьшить эти проблемы, необходимы эффективные системы обнаружения спама, и машинное обучение (ML), которое стало ключевым инструментом для создания таких систем. В данной работе рассматриваются два основных метода машинного обучения (ML) для обнаружения спама в электронной почте: сквозные методы глубокого обучения и традиционные методы, основанные на разработке функций.

Обзор темы

Обнаружение спама – это задача бинарной классификации, которая делит электронные письма на две категории: спам и не-спам. Чтобы преобразовать функции, основанные на традиционных методах, в общий формат, подходящий для моделей машинного обучения, необходимо вручную извлекать функции из содержимого электронной почты и метаданных. В этом методе часто используются такие модели, как наивный байесовский алгоритм [2], метод опорных векторов (SVM) и случайные леса.

Сквозные методы глубокого обучения используют нейронные сети для автоматического определения закономерностей и извлечения характеристик непосредственно из необработанных данных электронной почты. Это касается таких моделей, как трансформаторы (например, BERT), рекуррентные нейронные сети (RNNS) и сверточные нейронные сети (CNNS). С помощью этих моделей можно фиксировать более сложные связи и закономерности в данных, что может привести к повышению эффективности.

Актуальность темы, применимость в различных сферах

Системы обнаружения нежелательной почты необходимы:

- Поставщикам услуг электронной почты (ESP): Для улучшения взаимодействия с пользователями и защиты от вредоносного контента такие компании, как Google, Microsoft и Yandex, используют фильтры спама.
- Корпоративным почтовым системам: Компании используют средства обнаружения спама для защиты своих каналов связи, сохранения личных данных и предотвращения фишинговых атак [1].

- Интернет-провайдерам и компаниям, занимающимся сетевой безопасностью, они используют спам-фильтры для защиты пользователей от онлайн-опасностей и сохранения целостности сети.
- Пользователям персональных почтовых клиентов. Они получают преимущества от повышения безопасности электронной почты и уменьшения беспорядка в своих почтовых ящиках.

Предотвращая опасности, связанные со спамом, эффективные системы обнаружения спама повышают удобство работы пользователей, а также поддерживают более масштабные инициативы в области кибербезопасности.

Решение проблемы, связанной с обнаружением спама в электронной почте при помощи машинного обучения.

Сбор и подготовка данных

Надежный набор данных является основой любой системы обнаружения спама на основе машинного обучения (ML). Часто используются общедоступные наборы данных, такие как Ling-Spam, SpamAssassin Public Corpus и набор данных электронной почты Enron [3].

Этапы предварительной обработки данных:

- Очистка текста: удаление лишних пробелов, специальных символов и HTML-тегов.
- Нормализация: изменение различных сокращений и преобразование текста в нижний регистр.

Далее следует разделение текста на слова или лексемы с помощью разметки.

- Удаление стоп-слов: исключаются часто используемые термины, которые не помогают отличить спам.

- Лемматизация/Стемминг: сокращение слов до их основной формы.

Метод разработки функциональных возможностей

Разработка функциональных возможностей заключается в создании функциональных возможностей из необработанных данных электронной почты, чтобы их могли использовать обычные модели машинного обучения.

Методы извлечения функциональных возможностей:

- Набор слов (BoW): Этот метод представляет текст в виде набора значений слов.
- TF-IDF (Частота термина - обратная частота документа): Изменяет количество слов в соответствии с их значимостью во всем наборе данных.
- N-граммы: содержит контекст, фиксируя последовательности слов из n.
- В метаданные включаются характеристики электронной почты, такие как адрес отправителя, строка темы и статус вложения.
- Особенности содержимого: HTML-теги, пунктуация и частота использования определенных терминов.

Выбор модели и инструктаж:

- Наивная байесовская модель: вероятностная модель, которая предполагает независимость функций и применяет теорему Байеса.
- Метод опорных векторов (SVM): определяет гиперплоскость, которая эффективно разделяет электронные письма на категории спама и не-спама.

- Случайный лес: комплексный метод, который повышает точность классификации за счет использования нескольких деревьев решений.

Показатели для оценки:

- Точность: процент точно классифицированных электронных писем и процент действительно положительных результатов обнаружения спама среди всех положительных результатов.
- Количество отзывов: процент реальных спам-писем, которые обнаруживаются как действительно положительные.
- Показатель F1: среднее значение для запоминания и точности, которое уравнивает эти два показателя.

Комплексные методы глубокого обучения

Эти модели работают непосредственно с необработанными текстовыми данными, самостоятельно подбирая функции по мере обучения.

Представление данных:

- Встраивание слов: используется Word2Vec и GloVe для представления слов в виде плотных векторов в непрерывном пространстве.
- Контекстуальное встраивание: чтобы улучшить качество представления, используются такие модели, как BERT, для отображения значения слова в контексте.

Построение моделей:

- Сверточные нейронные сети (CNN): идентифицируются иерархические характеристики и локальные шаблоны в тексте.

- Рекуррентные нейронные сети (RNN): такие как LSTM и GRU, полезны для последовательных данных и могут фиксировать временные зависимости.
- Трансформаторы: сложные модели, такие как BERT, которые параллельно анализируют полные последовательности и фиксируют взаимозависимость на большом расстоянии, используя методы самоанализа.

Обучение и оценка:

- Используя наборы данных электронной почты с пометками, обучаются модели глубокого обучения.
- Применяются те же критерии для оценки: F1-оценка, отзывчивость и аккуратность.
- Учитываются дополнительные показатели, такие как интерпретируемость модели, вычислительные ресурсы и время обучения.

Сравнительная оценка

Следующие стандарты рассматриваются для того, чтобы сравнить разработку функций и комплексные методы:

1. Производительность:
   - Точность: поскольку сквозные модели могут фиксировать сложные закономерности, их точность, как правило, выше.
   - Точность и запоминаемость: модели глубокого обучения часто обладают более высокой точностью и запоминаемостью, что снижает количество ложных срабатываний и ложноотрицательных результатов.
2. Сложность:

- Время обучения: хотя традиционные модели обучаются быстрее, им может потребоваться много функциональных возможностей.
- Вычислительные ресурсы: память и вычислительная мощность являются основными требованиями к моделям глубокого обучения.

3. Масштабируемость:

- Обычные модели: лучше подходят для больших наборов данных, но их сложнее масштабировать с помощью выбора функций.
- Модели глубокого обучения: требуют эффективного аппаратного обеспечения, но более масштабируемы для работы с большими наборами данных.

4. Реализация:

- Простота использования: Традиционные подходы больше подходят для более простых приложений из-за их простоты реализации и интерпретации.
- Гибкость: Модели глубокого обучения более сложны, но обеспечивают большую гибкость при обучении на основе различных источников данных.

Вывод

Эмпирическая оценка показывает, что модели комплексного глубокого обучения лучше справляются с обнаружением спама, особенно те, которые используют сложные архитектуры, такие как BERT. Но это увеличивает количество времени и ресурсов, необходимых для обучения. Несмотря на то, что традиционные модели менее точны, их можно внедрять быстрее и они проще в использовании, что делает их подходящими для приложений с ограниченными ресурсами.

Данный обзор демонстрирует преимущества и недостатки сквозных и функционально-инженерных методов обнаружения нежелательной почты. Сквозные модели более надежны и точны, что делает их подходящими для ситуаций с высокими ставками, когда точность имеет решающее значение. Благодаря простоте использования и сниженным требованиям к ресурсам традиционные модели по-прежнему актуальны и являются приемлемым вариантом для небольших по масштабу применений. Будущие исследования могут быть сосредоточены на гибридных моделях, которые сочетают в себе лучшие характеристики двух методов для повышения эффективности обнаружения спама при одновременном снижении сложности и потребления ресурсов.

Использованные источники:

1. Бурмистров Д.А., Валикова М.А., Жукова М.Г. Методы создания антивирусных алгоритмов для распознавания спама в электронных сообщениях // Научно-технический вестник информационных технологий, механики и оптики. Volume 18, 2018, С. 113-120.
2. Зуев М.Ю. Детали алгоритмов машинного обучения для обнаружения спама в электронной почте // Сборник трудов конференции "Информационные технологии и системы". Volume 37, 2019, С. 134-139.
3. Н. Кумар, С. Соновал и Нишант, «Обнаружение спама по электронной почте с использованием алгоритмов машинного обучения», в материалах Второй международной конференции по вычислительным приложениям для изобретательских исследований (ICIRCA), стр. 108-113, IEEE, Коимбаторе, Индия, июль 2020 года.