

Кудакаев И.Ф., Ракипов Д.А.
Уфимский университет науки и технологий, Уфа

Научный руководитель:
Юнусова Д.С.
Уфимский университет науки и технологий, Уфа

ОБНАРУЖЕНИЕ СПАМА В EMAIL

Аннотация. В статье рассматриваются базовые подходы к обнаружению спама в электронной почте с применением методов обработки естественного языка (NLP). Пошаговое руководство ориентировано на начинающих специалистов и студентов, интересующихся машинным обучением и лингвистическим анализом текстов. Подробно описываются этапы подготовки данных, методы предобработки текста (включая лемматизацию и токенизацию), выбор признаков и алгоритмы классификации. Отдельное внимание уделяется сравнительному анализу моделей и оценке их эффективности. Работа сопровождается практическими рекомендациями и примерами кода, что делает ее полезной для создания собственного инструмента фильтрации спама.

Ключевые слова: обработка естественного языка, машинное обучение, фильтрация спама, классификация текста, лемматизация.

Обнаружение спама в электронной почте представляет собой одну из важнейших задач в области информационной безопасности и интеллектуального анализа текста. В связи с ростом объема нежелательной корреспонденции возрастает необходимость в автоматизированных решениях, способных эффективно фильтровать спам. Одним из наиболее результативных направлений в решении данной задачи является применение методов обработки естественного языка (Natural Language Processing, NLP), позволяющих анализировать текст писем, извлекать ключевые признаки и классифицировать сообщения по определенным критериям.

Использование электронной почты больше не ограничивается только общением с другим человеком. В эпоху цифровизации пользователь постоянно получает по электронной почте информацию о текущих предложениях, статусе заказа или новостях в социальных сетях. Инженерия знаний (Knowledge Engineering) и машинное обучение (Machine Learning) – это два основных подхода, которые используются при фильтрации электронной почты. В подходе инженерии знаний необходимо задать набор правил, согласно которым электронные письма классифицируются как спам (нежелательная почта) или хам (нормальная почта).

Такой набор правил должен быть создан либо самим пользователем фильтра, либо другим уполномоченным органом (например, компанией-разработчиком программного обеспечения, предлагающей инструмент фильтрации спама на основе правил). Применение этого метода не демонстрирует обнадеживающих результатов, так как правила необходимо постоянно обновлять и поддерживать, что требует времени и неудобно для большинства пользователей. Подход машинного обучения является более эффективным, чем подход инженерии знаний; он не требует задания правил [1]. Вместо этого используется набор обучающих примеров – в данном случае это заранее классифицированные электронные письма. Для успешного внедрения системы обнаружения спама начинающим специалистам рекомендуется использовать поэтапный подход: от сбора и подготовки данных до анализа результатов классификации.

Важно учитывать не только точность модели, но и такие метрики, как полнота, точность, F-мера, а также способность модели обрабатывать и распознавать сложные случаи, включая сарказм, скрытые намерения и измененные шаблоны спама. Таким образом, применение NLP в задаче фильтрации электронной почты требует комплексного подхода, сочетающего лингвистический анализ и математические методы классификации. Специальный алгоритм затем используется для извлечения правил классификации из этих писем. Подход машинного обучения широко изучен, и существует множество алгоритмов, которые могут быть применены для фильтрации электронной почты [2].

NLP может применяться как для лингвистического анализа текста, так и для анализа настроений и мнений (Sentiment Analysis), переводов, а также в голосовых помощниках, чат-ботах и системах «вопрос-ответ». В сочетании с аспектами информатики, такими как системный анализ и моделирование, а также с областями теоретической информатики, например, теориями вычислимости и сложности, компьютерная лингвистика получает знания о структурах данных и использовании эффективных алгоритмов. Это позволяет решать прикладные задачи, такие как перевод предложений на разные языки или обработка заказа пиццы по телефону [3].

Для достижения этих целей необходимо охватить множество задач – от распознавания устной речи до построения парсера и разработки набора методов и алгоритмов, с помощью которых естественный язык и человеческая коммуникация могут быть проанализированы в программных системах. При этом важно извлечь как можно более глубокое понимание устной или письменной речи. Основой для этого является источник текстовых данных, называемый корпусом. Поскольку корпуса и другие текстовые данные, как правило, представлены в неструктурированном виде, применяется широкий спектр методов, чтобы преобразовать необработанные данные в четко определенные последовательности лингвистических компонентов. Обработка естественного языка (Natural

Language Processing, NLP) занимается пониманием человеческой речи – и именно в этом заключается основная сложность.

Язык очень сложен и не всегда однозначен. Дополнительно анализ человеческой речи сильно усложняется такими явлениями, как фразеологизмы, ирония или игра слов. Недостаточно, чтобы компьютер просто распознавал отдельные слова. Важно, чтобы машина понимала предложения и их контексты. Человек в этом плане имеет преимущество: он может опираться на жизненный опыт и учит язык с рождения. Поэтому язык для человека понятен и легко поддается интерпретации. Компьютеру же необходимо освоить этот навык с нуля с помощью алгоритмов. Правила, по которым происходит анализ языка, могут быть разной степени сложности [4].

Разница между ними в том, что корень-основа (стем) не всегда является корректным словом с точки зрения лексикографии, то есть может отсутствовать в словаре. Лемма же всегда является словарной формой. Процесс лемматизации значительно медленнее, чем стемминг, поскольку требует дополнительного шага – определения леммы путем удаления аффикса только в том случае, если соответствующая лемма действительно существует в словаре. Это требует дополнительной мощности со стороны компьютерной лингвистики, например, использования теггера частей речи (Part-of-Speech-Tagger) [5].

Освоение базовых инструментов NLP и алгоритмов машинного обучения позволяет не только повысить точность фильтрации, но и развить практические навыки в области анализа текста. Для начинающих специалистов данная область открывает широкие возможности для экспериментов и дальнейших исследований, а также способствует созданию более безопасной и комфортной среды электронного общения.

Список использованных источников:

1. Горюнова Д.А., Кузьмина Е.С. Обнаружение спама в электронной почте на основе машинного обучения: сравнение функциональной разработки и комплексных методов // Экономика и социум. 2024. № 6-1 (121). URL: <https://cyberleninka.ru/article/n/obnaruzhenie-spama-v-elektronnoy-pochte-na-osnove-mashinnogo-obucheniya-sravnenie-funktsionalnoy-razrabotki-i-kompleksnyh-metodov-1> (дата обращения: 20.04.2025).

2. Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. Предобработка текстов электронных писем в задаче обнаружения спама // Труды учебных заведений связи. 2020. № 4. URL: <https://cyberleninka.ru/article/n/predobrabotka-tekstov-elektronnyh-pisem-v-zadache-obnaruzheniya-spama> (дата обращения: 20.04.2025).

3. Частикова В.А., Козачёк К.В. Обзор актуальных проблем основных методов фильтрации спама и анализ их эффективности // Вестник Адыгейского государственного университета. Серия 4: Естественно-

математические и технические науки. 2021. № 3 (286). URL: <https://cyberleninka.ru/article/n/obzor-aktualnyh-problem-osnovnyh-metodov-filtratsii-spama-i-analiz-ih-effektivnosti> (дата обращения: 20.04.2025).

4. Харитонов Ю.С., Савина В.С., Паньини Ф. Предвзятость алгоритмов искусственного интеллекта: вопросы этики и права // Вестник Пермского университета. Юридические науки. 2021. № 53. URL: <https://cyberleninka.ru/article/n/predvzyatost-algoritmov-iskusstvennogo-intellekta-voprosy-etiki-i-prava> (дата обращения: 20.04.2025).

5. Хаммасова Л.Ш., Нестеров С.А. Распознавание спам-сообщений с использованием методов машинного обучения // SAEC. 2023. № 3. URL: <https://cyberleninka.ru/article/n/raspoznvanie-spam-soobscheniy-s-ispolzovaniem-metodov-mashinnogo-obucheniya> (дата обращения: 20.04.2025).

Kudakaev I.F., Rakipov D.A.

Ufa University of Science and Technology, Ufa

Scientific supervisor:

Yunusova D.S.

Ufa University of Science and Technology, Ufa

EMAIL SPAM DETECTION

Abstract. This article explores fundamental approaches to email spam detection using natural language processing (NLP) techniques. Designed for beginners and students interested in machine learning and linguistic text analysis, the guide walks through the key stages of data preparation, text preprocessing (including lemmatization and tokenization), feature selection, and classification algorithms. Special focus is given to comparing models and evaluating their performance. The article includes practical tips and code examples, making it a valuable resource for those looking to build their own spam filtering tool.

Keywords: natural language processing, machine learning, spam filtering, text classification, lemmatization.