УДК 004.056.3:004.8

Огляд існуючих методів і алгоритмів виявлення та фільтрації спаму

Олександр Губар[0009-0005-2869-2373]
Валерій Козачок[0009-0005-2869-2373]

Київський столичний університет імені Бориса Грінченка, Україна

Анотація. У цій статті представлено систематичний огляд сучасних підходів до виявлення та фільтрації спаму в текстових та мультимедійних каналах комунікації. Проаналізовано класичні сигнатурні та статистичні методи, фільтри на основі чорних списків, а також методи машинного та глибинного навчання.
Ключові слова: спам фільтрація, наївний Баєс, машинне навчання, гібридні системи, масштабованість.

Review of Existing Methods and Algorithms for Spam Detection and Filtering

Oleksandr Hubar[0009-0005-2869-2373]
Valerii Kozachok[0009-0005-2869-2373]

Borys Grinchenko Kyiv Metropolitan University, Ukraine

Abstract. This article presents a systematic review of modern approaches to spam detection and filtering in text and multimedia communication channels. Classical signature-based and statistical methods, blacklist filters, as well as machine learning and deep learning techniques are analyzed.
Keywords: spam filtering, naive Bayes, machine learning, hybrid systems, scalability**.**

Вступ

Спам контент в онлайн середовищі зростає експоненційними темпами щодня. Оскільки такий контент водночас може приносити прибутки й завдавати значних збитків, ефективне запобігання його поширенню та якісний аналіз інцидентів стають критичною необхідністю. Класичні підходи вже не забезпечують належного рівня безпеки, оскільки спамери активно застосовують обхідні техніки, обфускацію та використання штучного інтелекту.

Методи та моделі

Основні методи виявлення та фільтрації спаму можна умовно поділити на кілька груп. Сигнатурні методи базуються на виявленні відомих шаблонів і ключових слів у повідомленнях, що забезпечує швидку і точну ідентифікацію раніше відомого спаму, проте є мало ефективними проти нових видів атак. Системи на кшталт SpamAssassin використовують понад 700 тестів для оцінки повідомлень, забезпечуючи високу швидкість обробки відомих зразків [1]. Фільтри на основі чорних списків використовують бази підозрілих адрес, що дає змогу оперативно блокувати спам із

відомих джерел з мінімальними витратами ресурсів, але не здатні розпізнати нові або змінені адреси відправників [2].

Статистичні методи, серед яких найпоширенішим є наївний байєсівський класифікатор, застосовують ймовірнісні моделі для оцінки спамності повідомлень, поєднуючи простоту реалізації із достатньою ефективністю до 95–97% виявлення спаму [3].

TF-IDF методи виділяють унікальні терміни в документах з точністю до 97.99%, але чутливі до підроблених розподілів слів. Методи машинного навчання (SVM, Random Forest, k-NN) підвищують точність класифікації, проте мають потребу у регулярному оновленні навчальних даних і моделей. Random Forest демонструє найвищу точність серед традиційних методів – до 99.91% з найнижчим рівнем помилкової класифікації (4.13%) [4–6].

Глибинне навчання, зокрема нейронні мережі типу CNN, RNN і трансформери, демонструє високу адаптивність і здатність працювати з великими обсягами текстових і мультимедійних даних, хоча й вимагає значних обчислювальних ресурсів. CNN ефективно обробляють текстові послідовності та виявляють локальні патерни, RNN та LSTM аналізують послідовні дані з точністю понад 97%, а BERT використовує механізм самоуваги та досягає точності 98.67%.

Гібридні підходи, які об'єднують кілька методів, дозволяють компенсувати недоліки окремих технологій і є найбільш перспективними для побудови сучасних систем фільтрації спаму. Комбінація CNN з RNN захоплює локальні та контекстні особливості, а ансамблеві методи використовують прогнози кількох моделей для підвищення надійності [4, 7].

Результати

Дослідження охопило порівняльний аналіз п’яти груп методів фільтрації спаму, серед яких сигнатурні, спискові, статистичні, традиційні алгоритми машинного навчання та методи глибинного навчання.

За результатами обробки експериментальних даних виявлено, що сигнатурні системи демонструють середню точність 90–94% при розпізнаванні відомих шаблонів повідомлень, у той час як продуктивність фільтрів на основі чорних списків залежить від частоти оновлення бази і знижується до 70–80% у разі надходження спаму з нових джерел, при цьому обидва підходи забезпечують високу пропускну здатність до 10 000 повідомлень/с і мінімальні обчислювальні витрати, проте виявляють обмежену адаптивність до лексичних і структурних модифікацій спам-контенту.

Статистичні алгоритми, зокрема наївний байєсівський класифікатор і TF-IDF, показали стабільну точність 95–97% у корпусах середнього обсягу, хоча їхня ефективність зменшується до 85–88% при аналізі мультимедійних або багатомовних даних, а в експериментальних випробуваннях Random Forest досягав точності 99,9% із показником хибнопозитивних спрацьовувань 4,13%, тоді як SVM забезпечував 98,5% при збереженні стабільності на великих обсягах даних; методи k-NN виявили чутливість до розміру вибірки і потребують попереднього нормування ознак для досягнення 93–95% точності класифікації.

Глибинні нейронні мережі продемонстрували найвищі результати, адже CNN забезпечували середню точність 97,5% для текстових послідовностей, RNN/LSTM – 96–97% за умови витрат у 2–3 рази більше часу на тренування, а трансформерна модель BERT досягала 98,67% завдяки двосторонньому контекстуальному аналізу. Тестування мультимодальних систем (текст + зображення) засвідчило зниження загальної точності на 3–4% через нерівномірність навчальних наборів.

Отримані результати дають змогу кількісно оцінити переваги й обмеження кожного підходу, що закладає основу для подальшої розробки гібридних антиспам-рішень.

Обговорення

Класичні методи фільтрації спаму, такі як сигнатурні аналізатори та чорні списки, забезпечують ефективний попередній відбір відомих загроз, але втрачають актуальність у разі складних чи мультимедійних атак через затримки оновлення сигнатур і можливість обходу зловмисниками. Статистичні підходи, зокрема наївний байєсівський класифікатор та TF-IDF, відзначаються простотою впровадження та швидкою обробкою тексту з точністю до 95–98%, втім їх ефективність падає при збільшенні обсягу мультимедійного й багатомовного спаму, що спричиняє зростання хибнопозитивних спрацьовувань і вимагає частого перенавчання моделей.

Алгоритми машинного та глибинного навчання демонструють найвищу точність понад 98%, і здатні розпізнавати складні патерни в тексті та мультимедійних даних, проте їх розгортання у масовому масштабі ускладнюють високі обчислювальні вимоги та відсутність прозорості, що створює труднощі з аудитом і поясненням рішень. Крім того, використання персональних даних для навчання моделей породжує етичні та правові ризики щодо конфіденційності користувачів, що обмежує можливості практичного застосування таких рішень.

Висновки

Проведений огляд методів виявлення та фільтрації спаму дозволив встановити, що класичні сигнатурні фільтри та системи на основі чорних списків залишаються ефективними інструментами початкової обробки повідомлень, проте їх точність різко знижується при зіткненні з новими, модифікованими або мультимедійними формами спаму. Статистичні алгоритми, зокрема наївний байєсівський класифікатор і TF-IDF, демонструють високу швидкодію та простоту реалізації, забезпечуючи 95–98% точності класифікації для текстових повідомлень, однак вони вразливі до змін у мовних патернах і часто генерують хибнопозитивні спрацьовування в умовах різноманітного мультимедійного та багатомовного спаму.

Алгоритми машинного навчання, такі як SVM, Random Forest і k-NN, забезпечують помітне підвищення точності класифікації завдяки здатності виявляти складні залежності в ознаках повідомлень. Зокрема, Random Forest демонструє до 99.9% точності з мінімальним рівнем помилкової класифікації, однак потребує великих навчальних даних і частого донавчання моделей для збереження актуальності результатів. Методи глибинного навчання, згорткові та рекурентні нейронні мережі, а також трансформери на кшталт BERT відзначаються високою адаптивністю до нових спам патернів і здатністю обробляти мультимедіа, але їх широке впровадження ускладнюється значними обчислювальними ресурсами та низькою інтерпретованістю прийнятих рішень.

Найбільш перспективним підходом сьогодні є інтеграція різних методологій в єдині гібридні системи, які поєднують переваги сигнатурного аналізу, статистичних моделей та методів машинного і глибинного навчання. Такі системи дозволяють досягати оптимального балансу між точністю, швидкодією та масштабованістю, знижувати кількість хибнопозитивних спрацьовувань і адаптуватися до швидко змінних загроз цифрового середовища. У подальших дослідженнях доцільно зосередитися на підвищенні прозорості алгоритмів, мінімізації ресурсних витрат і забезпеченні захисту конфіденційності користувацьких даних при навчанні моделей.

Джерела

1. Wang, X. (2024). Spam Filtering in the Modern Era: A Review of Machine Learning, Deep Learning, and System Comparisons. In Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (pp. 452–456). International Conference on Data Analysis and Machine Learning. SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0013526000004619

2. Yasin, A., & AbuAlrub, F. (2016). Enhance RFID Security Against Brute Force Attack Based on Password Strength and Markov Model. International Journal of Network Security & Its Applications, 8(5), (pp.19–38). https://doi.org/10.5121/ijnsa.2016.8402

3. Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

4. Kaddoura, S., Chandrasekaran, G., Elena Popescu, D., & Duraisamy, J. H. (2022). A systematic literature review on spam content detection and classification. PeerJ Computer Science, 8, e830. https://doi.org/10.7717/peerj-cs.830

5. Костюк, Ю., Довженко, Н., Мазур, Н., Складанний, П., & Рзаєва, С. (2025). Методика захисту Grid-середовища від шкідливого коду під час виконання обчислювальних завдань. Кібербезпека: освіта, наука, техніка, 3(27), 22–40. https://doi.org/10.28925/2663-4023.2025.27.710

6. Складанний , П., Костюк, Ю., Рзаєва, С., Самойленко, Ю., & Савченко, Т. (2025). Розробка модульних нейронних мереж для виявлення різних класів мережевих атак. Кібербезпека: освіта, наука, техніка, 3(27), 534–548. https://doi.org/10.28925/2663-4023.2025.27.772

7. Al-Kaabi, H., Darroudi, A. D., & Jasim, A. K. (2024). Survey of SMS Spam Detection Techniques: A Taxonomy. AlKadhim Journal for Computer Science, 2(4), (pp. 23–34). https://doi.org/10.61710/kjcs.v2i4.88