УДК 004.93
Мельник К.В., Мельник В.М., Маційбора А.С.
Луцький національний технічний університет

ДОСЛІДЖЕННЯ КОНТЕНТНИХ МЕТОДІВ РОЗПІЗНАВАННЯ СПАМУ

Мельник К.В., Мельник В.М., Маційбора А.С. Дослідження контентних методів розпізнавання спаму. В роботі надана загальна постанова методів розпізнавання спаму. Наведено огляд існуючих підходів до вирішення задачі розпізнавання спаму. Описані основні підходи, що використовуються в задачі розпізнавання спаму, визначено етапи процесу розпізнавання та розглянуті найбільш поширені математичні методи пошуку спаму. Розкрито особливості використання, переваги та недоліки зазначених методів. Зроблено висновок щодо необхідності подальшого розроблення алгоритмів розпізнавання на основі зазначених методів, що були б простими в реалізації, ефективними, мали низькі обчислювальні витрати при навчанні та високу якість класифікації в реальних завданнях.
Ключові слова: розпізнавання спаму, метод опорних векторів, штучні нейронні мережі.

Мельник К.В., Мельник В.М.. Мацибора А.С. Исследование контентных методов распознавания спама. В работе предоставлена общая постановление методов распознавания спама. Приведен обзор существующих подходов к решению задачи распознавания спама. Описаны основные подходы, используемые в задаче распознавания спама, определены этапы процесса распознавания и рассмотрены наиболее распространенные математические методы поиска спаму. Раскрыты особенности использования, преимущества и недостатки указанных методов. Сделан вывод о необходимости дальнейшей разработки алгоритмов распознавания на основе указанных методов, которые были бы простыми в реализации, эффективными, имели низкие вычислительные затраты при обучении и высокое качество классификации в реальных задачах.
Ключевые слова: распознавания спама, метод опорных векторов, искусственные нейронные сети.

Melnyk K.V., Melnyk V.M., Matsibora A.S. Exploration of content methods for spam detection. In this paper is given general resolution of the methods for spam detection. The existing methods review is given to solve the problem of spam detection. There are described some main approaches used for the task of spam recognition in this work with the steps determination on the recognition process and the most shared mathematical methods of spam searching. There are also revealed the using features with advantages and disadvantages of specified methods. It is made some conclusion about the further development necessity for recognition algorithms on the basis of these submitted methods that would been simply used in implementation, effective, with having low computing costs in the training process and high classification quality in the real tasks.
Keywords: spam detection, reference vector method, artificial neural network.

Актуальність та постановка проблеми. Спам у електронній пошті залишається проблемою, за даними [1], близько 60% всіх електронних листів є спамом. Хоча кількість небажаної кореспонденції зменшилась, але вона стала більш шкідливою. За допомогою спам листів поширюють не тільки рекламу, а й шкідливе програмне забезпечення, програми для вимагання грошей такі як Petya, Cryakl і Shade.

За останні роки було винайдено чимало способів боротьби з небажаною кореспонденцією. Нажаль, зловмисники стежать за протидією поширенню спаму й винаходять все нові прийоми для обходу фільтрів. Таким чином, всі дослідження в області боротьби з не запитуваною кореспонденцією надзвичайно актуальні в цей час.

Аналіз останніх досліджень та публікацій. Аналіз, розробку методів розпізнавання спаму досліджено в роботах А.М. Мироненка, П. Грехама, Викас П. Дешпанде, І. Терейковський [2–5]. Кожен із методів має свої переваги та недоліки, для практичного застосування необхідно виконати їх порівняння за функціональністю, сферою застосування, ефективністю тощо.

Мета дослідження – огляд основних методів розпізнавання спаму в тексті, розкрити їхні особливості використання, переваги та недоліки.

Виклад основного матеріалу дослідження та обґрунтування отриманих результатів. Спам — масова розсилка кореспонденції рекламного чи іншого характеру людям, які не висловили бажання її одержувати. Передусім термін «спам» стосується рекламних електронних листів.

Існує два основних підходи до боротьби з небажаною кореспонденцією, а саме визначення відправника як спамера, або аналізу листа, на основі чого робиться висновок, що він є спамом. Найбільш ефективним є комплексних захист (рис. 1) [6], що складається з наступних етапів: аналіз відправника; використання фільтрів; аналіз змісту листа.

Для аналізу змісту в листах застосовуються різноманітні методи, кожен з яких має свої переваги і особливості використання. Переважна більшість методів розпізнавання спаму так чи інакше засновані на припущені, що текст, який відносяться до спаму, має однакові ознаки (слова чи

словосполучення), і наявність чи відсутність таких ознак в тексті визначає його приналежність чи неприналежність до тієї чи іншої категорії.

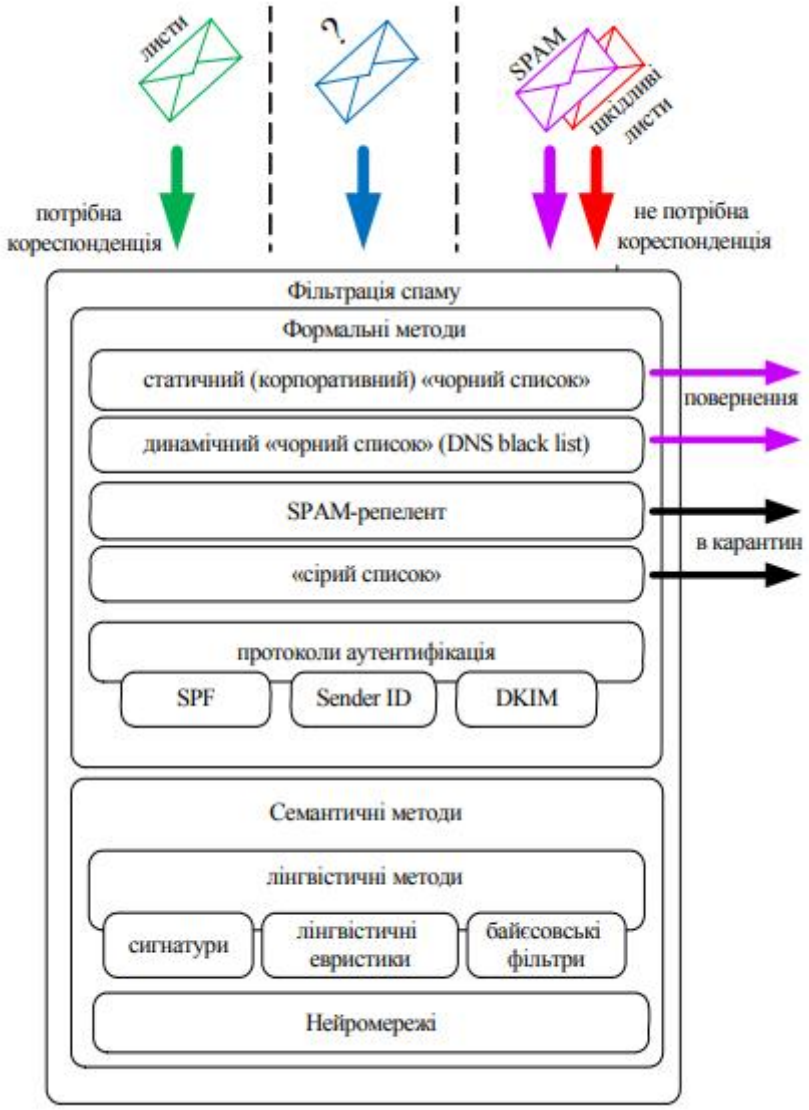

Рисунок 1. – Комплексний захист від спаму

Таким чином, задача фільтрації спаму, розглядається як задача класифікації – визначення належності об'єкта (електронного повідомлення) до одного з заздалегідь виділених класів (спам і «не спам») на підставі аналізу сукупності ознак, що характеризують даний об'єкт.

В даній статі описано та досліджено такі методи контекстного розпізнавання спаму в електронній пошті :

- Байесівська (наївна) класифікація спаму;
- класифікація методом опорних векторів;
- розпізнавання за допомогою штучних нейронних мереж.

Для порівняння даних методів використовувалась база електронних листів розміщена на сайті UCI Machine Learning Repository [7].

З усієї електронної кореспонденції було виділено 57 ознак, що характеризують дані повідомлення, а 58 змінна визначає приналежність листа до класу спам чи «не спам».

Дана база налічує 4601 спостереження з яких 1813 (39,4%) – це спам, а 2788 (60,6%) – звичайні листи.

Теорема Байєса лежить в основі багатьох сучасних систем штучного інтелекту, призначених для роботи в умовах невизначеності. Такі системи дають ймовірнісну оцінку, тому звичайно не

заміняють експерта, а забезпечують підтримку прийняття рішення. Таким чином, для спаму повинна бути визначена множина ознак.

У загальному вигляді визначення найбільш вірогідного класу алгоритмом наївною байєсівської класифікації виглядає наступним чином:

Нехай $F_S(W_i)$ – кількість спам-листів, у яких зустрілося слово (W_i), а $F_{NS}(W_i)$ – кількість корисних листів, у яких зустрілося слово W_i; H_S – гіпотеза про те, що лист є спамом, H_{NS} – корисний лист. Тоді ймовірність того, що поява слова W_i у листі означає спам, обчислюється за формулою:

$$P(W_i|H_S) = \frac{F_S(W_i)}{F_S(W_i)+F_{NS}(W_i)},$$

а ймовірність того, що слово W_i не вказує на спам у листі:

$$P(W_i|H_{NS}) = \frac{F_{NS}(W_i)}{F_S(W_i)+F_{NS}(W_i)},$$

Якщо вектор W включає всі m слів нового листа, то ймовірність того, що він спам, обчислюється за формулою Байєса таким чином:

$$P(H_S|W) = \frac{\prod_{j=1}^{m} P(W_j|H_S)}{\prod_{j=1}^{m} P(W_j|H_S)+\prod_{j=1}^{m} P(W_j|H_{NS})}.$$

Віднесення листа до спаму або корисних листів виконується з врахуванням заданого програмістом, адміністратором, користувачем почтової програми спам-фільтрації значення ймовірності, яке становить 0,6–0,8.

В основі фільтра лежить список ознак, за якими проводиться аналіз повідомлення і обчислюється умовна ймовірність спамності за кожного ознакою. Загальна ймовірність спаму повідомлення визначається методом об'єднання всіх ймовірностей за теоремою Байєса.

Провівши класифікацію даної бази листів за алгоритмом наївною байєсівської класифікації було складено звіт таб. 1. З результатів дослідження ми бачимо, що даний спосіб помилково фільтрує 767 (16,67 % від усіх спостережень) звичайних листів, як спам і дає всього 81,4% правильної фільтрації повідомлень.

Таблиця 1. Звіт розпізнавання спаму алгоритмом наївною байєсівської класифікації

Назва класу	Загалом	Правильно	Помилково	Правильно %	Помилково %
«Не спам»	2110	2021	89	95,78	4,22
Спам	2491	1724	767	69,21	30,79
Всього	4601	3745	856	81,4	18,6

Основні переваги наївного байєсівського класифікатора простота реалізації і низькі обчислювальні витрати при навчанні та класифікації. У тих рідкісних випадках, коли ознаки дійсно незалежні (або майже незалежні), наївний байєсівський класифікатор (майже) оптимальний. Основним недоліком методу є відносно невисока якість класифікації в більшості реальних завдань. Зазначений метод часто використовується в якості базового методу при порівнянні різних методів машинного навчання.

Метод опорних векторів (Support Vector Machine, SVM) використовує процес пошуку площини вирішення, яка може розділити позитивні і негативні приклади в багатовимірному просторі функції, в якому навчальні документи представлені як вектори. Цей метод розроблений В. Вапником в 1995 році, був вперше застосований до задачі класифікації текстів Торстеном Джохімсом. У своєму первинному вигляді алгоритм вирішував завдання розрізнення об'єктів двох класів. Метод набув величезну популярність завдяки своїй високій ефективності. Багато дослідників використовують його в роботах, присвячених класифікації текстів. Підхід, запропонований В. Вапником для визначення того, до якого з двох заздалегідь визначених класів повинен належати аналізований зразок, заснований на принципі структурної мінімізації ризику.

Результати класифікації текстів за допомогою методу опорних векторів є одними з найкращих, у порівнянні з іншими методами машинного навчання. Однак, швидкість навчання даного алгоритму одна з найнижчих.

Провівши класифікацію даної бази листів за алгоритмом методом опорних векторів було складено звіт таб. 2. З результатів дослідження ми бачимо, що даний спосіб дає 89,18% правильної фільтрації повідомлень. Даний метод працює краще ніж алгоритм наївною байєсівської класифікації.

Таблиця 2. Звіт розпізнавання спаму методом опорних векторів

Назва класу	Загалом	Правильно	Помилково	Правильно %	Помилково %
«Не спам»	2788	2645	143	94,87	5,13
Спам	1813	1458	355	80,42	19,58
Всього	4601	4103	498	89,18	10,82

Штучні нейронні мережі (Artificial Neural Network). Штучні нейронні мережі набули широкого вивчення в галузі аналізу даних з 1986 року. Вони представляють собою математичну модель, а також її програмні або апаратні реалізації, побудовані за подібністю мереж нервових клітин живого організму. Нейронні мережі - це один з найбільш відомих і старих методів машинного навчання.

Штучні нейронні мережі - це адаптивна система, яка складається з групи з'єднаних штучних нейронів. Система може бути навчена для зміни її внутрішніх станів, відображення зв'язків повідомлень та їх класів. Для ефективного проведення фільтрації листів необхідно визначити раціональну структуру і топологію нейронної мережі. Основні топології класифікуючих нейронних мереж - це одно- і багатошаровий персептрон, мережа радіально базисних функцій. Всі вищевказані топології мають високу точність в обробці одночасно лінійних і нелінійних прикладів, але прийняття рішень щодо класифікації важко формалізуються у зв'язку з природою організації нейронної мережі та представляють нетривіальну задачу з урахуванням масштабованості з обмеженими обчислювальними ресурсами.

Для дослідження було побудовано декілька різних нейронних мереж таб. 3.

Таблиця 3. Опис нейронних мереж

№	Архітектура	Продуктив-ність	Алгоритм	Функція помилки	Активація прихованих нейронів	Активація вихідних нейронів
1	MLP 57-19-2	96,43	BFGS 175	Сума квадратів	Експонента	Тотожна
2	MLP 57-24-2	95,34	BFGS 75	Крос-ентропія	Експонента	Софтмак
3	MLP 57-8-2	94,47	BFGS 53	Крос-ентропія	Експонента	Софтмак
4	MLP 57-8-2	96,21	BFGS 71	Сума квадратів	Гіперболічна	Логістична
5	RBF 57-30-2	60,53	RBFT	Сума квадратів	Гауссіан	Тотожна

Найкращі результати з них показали нейроні мережі № 1, № 4 та № 2.

Таблиця 4. Звіт розпізнавання спаму MLP 57-19-2 алгоритм BFGS 175

Назва класу	Загалом	Правильно	Помилково	Правильно %	Помилково %
«Не спам»	2788	2697	91	96,74	3,26
Спам	1813	1711	102	94,37	5,63
Всього	4601	4408	193	95,81	4,19

Таблиця 5. Звіт розпізнавання спаму MLP 57-8-2 алгоритм BFGS 71

Назва класу	Загалом	Правильно	Помилково	Правильно %	Помилково %
«Не спам»	2788	2716	72	97,42	2,58
Спам	1813	1690	123	93,27	6,73
Всього	4601	4406	195	95,78	4,23

Таблиця 6. Звіт розпізнавання спаму MLP 57-24-2 алгоритм BFGS 75

Назва класу	Загалом	Правильно	Помилково	Правильно %	Помилково %
«Не спам»	2788	2688	100	96,41	3,59
Спам	1813	1684	129	92,89	7,12
Всього	4601	4372	229	95,02	4,98

Отже, ми бачимо що нейроні мережі дають найкращі результати, тому це перспективний напрямок подальшого дослідження. Поєднання двох різних архітектур нейронних мереж може дати дуже продуктивні результати.

Висновки та перспективи подальшого дослідження. В результаті виконаного дослідження було проаналізовано переваги та недоліки таких методів розпізнавання спаму в електронних листах: Байесівська (наївна) класифікація спаму, класифікація методом опорних векторів та розпізнавання за допомогою штучних нейронних мереж. Було побудовано декілька різних нейронних мереж та порівняно їхню продуктивність з іншими способами фільтрації електронної кореспонденції.

В майбутньому дані проведеного дослідження можна використати для розробки спам-фільтра для електронної скриньки кафедри. Перспективним є дослідження продуктивності різних топології нейронних мереж для виконання даного завдання.

1. traffic from January 2014 to September 2017, by month [Електронний ресурс] — Режим доступу: https://www.statista.com/statistics/420391/spam-email-traffic-share/
2. Vikas P. Deshpand. An Evaluation of Naive Bayesian Anti-Spam Filtering Techniques / Vikas P. Deshpande, Robert F. Erbacher, Chris Harris // Proceedings of the 2007 IEEE Workshop o n Information Assurance United States Military Academy, West Point, 2007. — NY 20—22 June. — Режим доступу: http://digital.cs.usu.edu/~erbacher/publications/Bayes-Vikas2.pdf.
3. Graham P. A Plan for Spam / P. Graham, 2002. [Електронний ресурс] — Режим доступу: http://www.paulgraham.com/spam.html.
4. Мироненко А. Н. Алгоритм контентной фильтрации спама на базе совмещения метода опорных векторов и нейронныхсетей : автореф. дис. на соискание науч. степени канд. техн. наук: спец. 05.13.19 «Методы и системы защиты информации, информационная безопасность» / А. Н. Мироненко. — СПб., 2012. — 18 с.
5. Терейковський І. Методологія класифікації листів електронної пошти з використанням нейроних мереж. / І. Терейковський // Захист інформації – 2013, – Том 15, №2, – С. 115-122.
6. Кузьма К., Зівенко В. Аналіз методів фільтрації електронної пошти від спаму. / К. Кузьма, В. Зівенко // Науковий журнал – Геометричне моделювання та інформаційні технології № 1 (3), квітень 2017. – Миколаїв : МНУ імені В. О. Сухомлинського, 2017. — С. 84-89.
7. Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.