DOI 10.15276/imms.v15.no2.247
УДК 004.056.55

Informatics and Mathematical Methods in Simulation
Vol.15 (2025), No. 2, pp. 247-259

АЛГОРИТМ ВИЯВЛЕННЯ ФІШИНГУ В ЛИСТАХ МЕСЕНДЖЕРІВ ТА ЕЛЕКТРОННОЇ ПОШТИ ІЗ ВИКОРИСТАННЯМ НЕЙРОННИХ МЕРЕЖ З ФІКСОВАНОЮ КІЛЬКІСТЮ РАНЖОВАНИХ ЧИННИКІВ РИЗИКУ

В.О. Назаров, А.В. Садченко, О.А. Кушніренко

Національний університет «Одеська політехніка»
1, Шевченка пр., м.Одеса, 65044, Україна
Emails: koa@op.edu.ua, sadchenko@op.edu.ua

Стаття присвячена розробці алгоритмів виявлення найбільш шкідливого різновиду спаму – фішингу на основі нейронних мереж із сигмоїдною та пороговою функціями активації. Проведено огляд та аналіз існуючих підходів до виявлення спаму і встановлено, що в більшості сучасних алгоритмів використовується кількісний підхід заснований на постійному поповненню словників словосполучень у явному та токенізованому вигляді, що при використанні, наприклад, нейромережі на етапі прийняття рішення може збільшувати імовірність віднесення етичного змісту повідомлень та листів до спаму. Представлено новий підхід до формування вектору вхідних чинників, що можуть бути присутні як у спамі так і в етичному контенті. Було виділено чотири основних загальних чинника, що ранжуються по десятибальної шкалі ризику, це «можлива винагорода», «обмеження у часі», «емоційне забарвлення» та «наявність посилання на зовнішній ресурс». Щодо перших 3-х чинників запропонована 10-ти бальна шкала оцінки впливу згідно аналізу наявності шаблонів-словосполучень, а «наявність посилання на зовнішній ресурс» має дві градації: «1», якщо посилання є, та «0» якщо його немає. Таке ранжування дозволило створити цифрове уявлення факторів ризику у вигляді вхідного вектору чинників, що потім подається на вхід нейромережі із сигмоїдною чи пороговою функціями активації. Нейромережа із використанням сигмоїдної функції активації після навчання, на основі десяти випадків фішинг контенту та такої же кількості етичних листів, забезпечила імовірність виявлення шкідливого контенту приблизно 10^{-3} при умові віднесення підозрілих листів до категорії спам. Нейромережа із пороговою функцією активації і додатковим розбиттям діапазону чинника «наявність посилання на зовнішній ресурс» на десять умовних градації, за допомогою аналізу S матриці взаємного впливу, показала високу швидкість навчання та імовірність виявлення шкідливого контенту приблизно $0.5 \cdot 10^{-3}$. Результати виконаних досліджень можна використовувати при створені фішингових фільтрів як із жорстким, так і з м'яким рішенням, що призначені для вбудовування в поштові сервіси та месенджери соціальних мереж.

Ключові слова: фільтри спаму, фішингові листи, нейромережа із сигмоїдною функцією активації, чинники ризику.

Вступ. В сучасних реаліях дистанційна освіта та робота набувають вимушеної популярності. І якщо ще зовсім недавно основним каналом дистанційного обміну інформації була електронна пошта [1], то на сьогоднішній день оперативний обмін інформацією здійснюється через численні месенджери, такі як Viber, Telegram, WhatsApp. При цьому поштові сервіси, наприклад, від Google містять потужні інструменти щодо боротьби зі шкідливими чи спам-листами. Спам-фільтри зазвичай інтегровані до поштових сервісів і побудовані із використанням технологій на базі штучного інтелекту та нейронних мереж [2,3]. У таких умовах для успішної, з погляду зловмисника, доставки шкідливої інформації необхідно застосовувати дедалі складніші методи обходу [4,5] спам-захисту. Наприклад, адреса відправника може бути підроблена під адресу корпоративної пошти [6].

На відміну від самонавчених технологій поштових сервісів, месенджери в більшості випадків такого захисту не мають або використовують пропрієтарні

алгоритми з нерегламентованим ступенем захисту. Тому актуальним є завдання пошуку нових підходів до виявлення шкідливої інформації.

Під шкідливою інформацією чи спам-повідомленнями будемо мати на увазі такі повідомлення [7], що надіслані великій групі одержувачів без їх попередньої згоди.

Серед спам-листів є відносно безпечні, такі що використовуються для реклами товарів чи послуг. Більшою небезпекою володіють листи чи повідомлення що містять посилання [8] на веб-сайти зловмисного програмного забезпечення або фішингового хостингу, через які може бути вкрадено приватна інформація чи безпосередньо гроші. При розповсюдженні небезпечних повідомлень за допомогою месенджерів може використовуватись ID відправника, що добре відомий отримувачу.

Розглянемо основні підходи до реалізації спам-фільтрів.

Спам-фільтри на основі списків. Метод боротьби зі спамом за допомогою спам-фільтрів на основі списків [9,10] передбачає використання:

– чорних списків (blacklists) – список адрес/доменів, з яких ніколи не слід приймати повідомлення;

– білих списків (whitelists) – список надійних відправників, чиї листи завжди пропускаються;

– сірих списків (greylists) – тимчасово відхиляються листи від невідомих відправників, з можливістю їх повторної доставки пізніше.

Переваги спам-фільтрів на основі списків:

– простота реалізації та легкість налаштування, зрозумілість у використанні;

– висока ефективність проти відомих джерел спаму – якщо відправник вже у чорному списку, його листи будуть заблоковані;

– низькі обчислювальні витрати що не вимагають складного аналізу тексту або машинного навчання;

– мінімальна ймовірність помилкових спрацьовувань щодо білих списків та гарантована доставка від надійних відправників.

Недоліки спам-фільтрів на основі списків:

– обмежена гнучкість – не здатні справлятися з новим або змінним спамом (невідомі джерела);

– високий ризик хибно позитивних спрацьовувань – легітимний відправник може опинитися в чорному списку помилково;

– залежність від оновлення списків – потрібне постійне оновлення для збереження актуальності;

– можна обійти – спамери легко змінюють адреси або використовують зламані сервери, щоб не потрапити до списку;

– сірі списки можуть викликати затримки доставки, особливо для нових відправників.

Байєсівські фільтри. Принцип дії байєсовського фільтру [11-16] заснований на статистичному аналізі тексту листа на основі ймовірності того, що він містить спам, тобто використовується модель навчання із вчителем.

Серед переваг байєсовських фільтрів можна виділити можливість самонавчання, ефективність, особливо при великому обсязі попередньо оброблених даних. Такі фільтри здатні виявляти новий спам, навіть із невідомих адрес.

До недоліків байєсовських фільтрів можна віднести наступне:

– потрібність попереднього навчання на великій вибірці спаму та нормальних листів;

– складність визначення стану навчання завдяки чому можлива помилкова ідентифікація шкідливого контексту;

– можливе різке погіршення якості роботи фільтру при використанні методів обфускації (маскування) тексту [17].

Евристичні фільтри. Принцип дії евристичного фільтру [18] полягає в формуванні спеціалізованого набору правил і шаблонів щодо змісту листів, наприклад, наявність підозрілих посилань, певних слів, формату листа.

До переваг евристичних фільтрів можна віднести:

– гнучкість завдяки доброму виявлянню шаблонного спаму;

– відсутність залежності від конкретного відправника.

Недоліки евристичних фільтрів:

– можуть давати помилкові спрацьовування при відсутності відповідного шаблону;

– вимагають постійного поповнення словнику ключових слів;

– важко підтримувати універсальність правил.

Використання нейромережі. Переваги нейромереж (НМ) [19-21] у виявленні спаму:

– краще відношення вірно виявлених випадків спаму до загального числу проаналізованих повідомлень порівняно із іншими методами;

– нейромережі можуть навчатися з нових даних і адаптивно підлаштовуватися під зміну тактик зловмисників;

– нейромережі можна використовувати щодо виявлення складних шаблонів текстів, на відміну від простих фільтрів, наприклад, на основі аналізу ключових слів, нейромережі виявляють складні взаємозв’язки між словами, стилем, структурою повідомлень тощо.

До недоліків існуючих алгоритмів виявлення спаму із використанням нейромережі можна віднести:

– необхідність великої кількості заздалегідь підготовлених даних, що мають властивість у вигляді спам чи не спам;

– високі вимоги до обчислювальних ресурсів, таких як навчання та обробка великих обсягів даних, наявності потужних процесорів чи графічних карт та оперативної пам’яті;

– наявність ефекту «Чорна скринька» – іноді важко зрозуміти, чому мережа класифікувала лист як спам;

– наявність ефекту «Перенавчання» – якщо не ввести обмеження на кількість листів щодо навчання, може відбутись поступове погіршення характеристик виявлення шкідливого контенту;

– уразливість до атак – нейромережі можуть бути обмануті (наприклад, додаванням "етичних" словосполучень у лист, щоб обійти фільтр).

Загальним для вище перелічених методів виявлення шкідливого контенту в повідомленнях є відсутність критеріїв щодо ранжування рівня загрози [22-27], що виходить від окремих виявлених чинників та взаємного впливу дієвих чинників листу на його кінцеве призначення.

Метою даної статті є розробка алгоритму виявлення фішинг-спаму на основі одношарової нейромережі, та методу ранжування рівнів загрози, що виходить від спільних чинників щодо етичних та спам листів.

Основна частина. Розглянемо основні етапи аналізу листів з метою виявлення шкідливого контенту, рис 1.

Одним з найскладніших етапів даного алгоритму є виявлення шкідливих рис вмісту листів, що аналізуються, чи чинників впливу, ранжування по ступеню риску та перетворення у числову форму для формування вхідного вектору нейромережі.

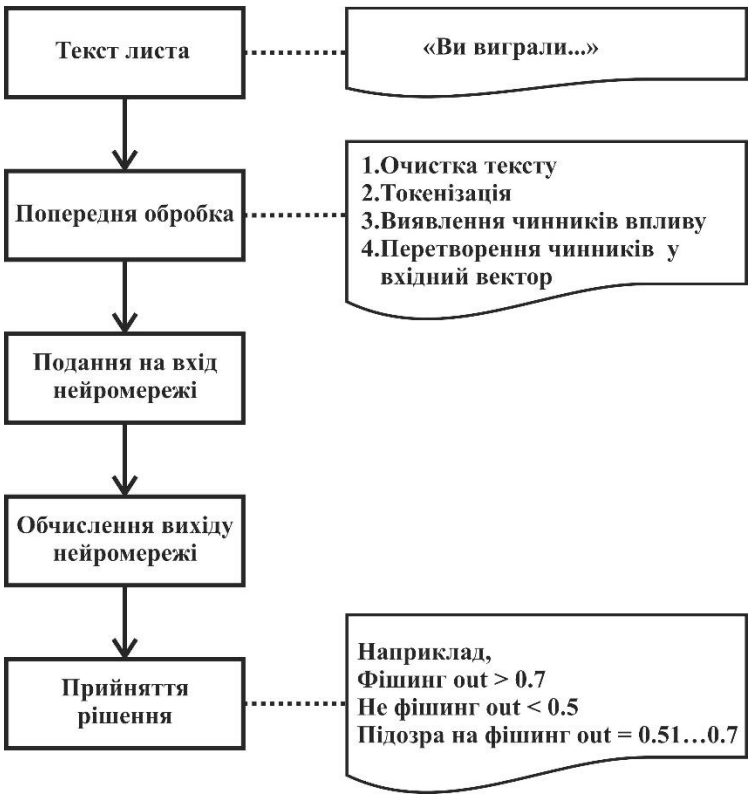

Рис. 1. Загальна структура алгоритму побудови антіспам-фільтру

Для виявлення найбільш впливових чинників ризику було проаналізовано зміст найбільш поширених фішингових листів, що стосуються отримання грантів, виграшу у лотереї та банківської діяльності.

Загальні риси шкідливих листів. Аналіз найбільш шкідливих спам-листів, тобто таких, що можуть призвести до фінансових та іміджевих втрат дозволяє виявити основні чинники, що можуть використовуватись для здійснення маніпуляції:

- психологічний тиск;
- шантаж або маніпуляція;
- запити на конфіденційні дані;
- підроблені або масковані посилання;
- маніпулятивна мова;
- відсутність чіткої ідентифікації відправника;
- відсутність альтернатив зв’язку;
- обмеження у часі;
- обіцянки великих виграшів або ексклюзивних можливостей;
- емоційне забарвлення та перебільшення.

Слід зазначити, що листи які уявляють із себе спам-фішинг завжди містять посилання на зовнішній шкідливий ресурс.

Однак е загальні чинники, що можуть міститись як в етичних, так і в спам-листах:

- запити на конфіденційні дані;
- відсутність чіткої ідентифікації відправника;
- відсутність альтернатив зв’язку;
- обмеження у часі;
- обіцянки великих виграшів чи винагорода за виконану роботу;
- емоційне забарвлення та перебільшення.

З метою створення математичної моделі алгоритму виявлення спаму пропонується створити *вектор чинників*, що є загальними для етичних та спам-листів.

Визначення балів щодо окремих складових можливе завдяки використанню нейромережі з метою періодичного оновлення вектору чинників.

Виявимо чинники, що можливо уявити у вигляді числа балів. Наприклад, деякі медичні застосунки пропонують ліки в залежності від суб'єктивної оцінки самопочуття в балах.

Розглянемо декілька текстів спам-листів, що містять вказівку суми нагороди у явному вигляді.

Спам-лист, що містить інформацію про фіктивний грантовий проект: «Ваша успішність перевершує очікування, і Вам надана можливість отримати грантову підтримку у розмірі 100 000 USD для розвитку Ваших проектів».

Спам-лист, що містить інформацію про виграш в лотерею: «Ми з величезним задоволенням повідомляємо, що Ваша участь у престижній лотереї Royal Fortune Jackpot була успішною, і Ви стаєте щасливим переможцем головного призу у розмірі 75 000 USD!»

Ще варіант спам-листу, що містить інформацію про виграш в лотерею:

«Це надзвичайно важливе повідомлення! Ми з великим захопленням і водночас із глибокою тривогою повідомляємо, що результати останньої лотереї Platinum Prize Bonanza підтвердили Ваш виграш у розмірі 50 000 USD!»

Позначимо параметр миттєвої винагороди чи плати за виконану роботу як R_{m}. Максимальний запропонований виграш згідно проаналізованих спам-листів складає 100 000 USD, тому, щоб отримати оцінку R_{m} по 10-ти бальної шкалі:

$$R_m, балів = \begin{cases} 10, & R_m > 100000 \\ 2\left[Log_{10} R_m\right], & R_m < 100000 \end{cases},$$

де $\left[X\right]$ – ціла частина числа.

Параметр обмеження у часі T_{m} також приведемо до системи оцінювання за 10-ти бальної шкалою.

Приклад відповідності змісту листів чиннику обмеження у часі приведений в таблиці 1. Відсутність обмежень означає що $T_{\text{m}} = 0$.

Таблиця 1.

Відповідність змісту листів чиннику обмеження у часі T_{m}

Зміст листа	$\boldsymbol{T_{\text{m}}}$**, балів**
обмеження відсутні	0
«…до визначеної дати»	1
«…протягом двох тижнів»	2
«…протягом тижня»	3
«…бажано до післязавтра»	4
«…бажано до завтра.»	5
«…протягом наступних 24 годин»	6
«…протягом наступних 12 годин»	7
«Ви маєте лише 6 годин»	8
«…будь-яка затримка»	9
«Вам необхідно негайно»	10

Параметр емоційного забарвлення та перебільшення позначимо E_{z}.

Приклад відповідності змісту листів чиннику емоційного забарвлення приведений в таблиці 2. Відсутність забарвлення означає що $E_z = 0$.

Таблиця 2.

Відповідність змісту листів чиннику емоційного забарвлення

Зміст листа	E_z**, балів**
відсутність забарвлення	0
оплата виконаної роботи гарантована	1
ми цінуємо Ваш час тому буде надано авансовий платіж	2
при вчасному виконанні роботи розмір винагороди буде збільшено	3
Вам надана можливість отримати грантову підтримку	4
Вам нараховано бонуси	5
Ваша успішність перевершує очікування	6
розмір винагороди було збільшено	7
Ви претендуєте на отримання премії	8
Ви стаєте щасливим переможцем головного призу	9
нагороду вже надіслано	10

Також введемо чинник ризику R_z, що буде відповідати за імовірність хибної ідентифікації етичного листа як спам-листу. Також цей чинник необхідно прив'язати до факту наявності посилання на зовнішній шкідливий ресурс, наприклад якщо посилання нема то початкове значення $R_z = 0$, а якщо посилання є то $R_z = R_{z\,max}$.

Отримаємо наступний загальний вигляд вектору чинників, що містяться в листі чи повідомленні:

$$V_p = \left[R_m, T_m, E_z, R_z\right],$$

де R_m – чинник винагороди;

T_m – чинник обмеження у часі;

E_z – чинник емоційного забарвлення;

R_z – чинник ризику.

Опишемо алгоритм визначення спам-листів із використанням одношарової нейромережі з використанням нейрону з найбільш поширеною – сигмоїдною функцією активації та критерієм середньоквадратичного відхилення щодо отримання помилки. Архітектура нейрону приведена рис 2.

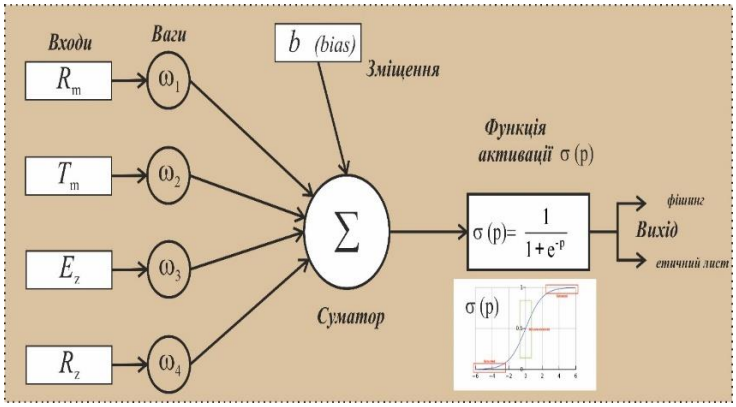

Рис. 2. Нейрон з 4-мя входами та сигмоїдною функцією активації

Крок 1. Обчислюємо значення аргументу p функції активації із використанням виразу:

$$p = V_p{}^T W + b = \sum_{i=1}^{4} v_i w_i + b = R_m w_1 + T_m w_2 + E_z w_3 + R_z w_4 + b,$$

де ω_1, ω_2, ω_3, ω_4 – це вагові коефіцієнти,

b – значення параметру початкового зсуву (bias).

Обчислюємо вихідне значення нейрону:

$$y = \sigma(p) = \frac{1}{1+e^{-p}}$$

Крок 2. Обчислюємо значення похибки чи параметр втрат L як середньоквадратичне відхилення (MSE) отриманого значення y від цільового значення y_a:

$$L = \frac{1}{2}(y - y_a)^2$$

Крок 3. Знайдемо похідну функції помилки з використанням градієнтного методу чи методу зворотного розповсюдження помилки:

$$\frac{dL}{dw_i} = \frac{dL}{dy} \cdot \frac{dy}{dp} \cdot \frac{dp}{dw_i},\ i = \overline{1,4}$$

Похідна помилки по вихідному значенню нейрона:

$$\frac{dL}{dy} = \frac{d}{dy}\left(\frac{1}{2}(y - y_a)^2\right) = 2 \cdot \frac{1}{2}(y - y_a) = y - y_a$$

Похідна вихідного значення по аргументу функції активації нейрона:

$$\frac{dy}{dp} = \frac{d}{dp}\left(\frac{1}{1+e^{-p}}\right) = -\frac{\left(1+e^{-p}\right)_p'}{\left(1+e^{-p}\right)^2} = -\frac{(1)_p' + (-z)_p' \cdot e^{-z}}{\left(1+e^{-p}\right)^2} = -\frac{0 + (-1 \cdot (z)_p') \cdot e^{-p}}{\left(1+e^{-p}\right)^2} = \frac{e^{-p}}{\left(1+e^{-p}\right)^2}$$

Похідна помилки щодо кожної ваги та параметра b:

$$\frac{dL}{dw_i} = \frac{dL}{dp} \cdot v_i,\ i = \overline{1,4}$$

$$\frac{dL}{db} = \frac{dL}{dp} \cdot b,$$

де $\frac{dL}{dp} = \frac{dL}{dy}\frac{dy}{dp}$

Крок 4. Виконуємо оновлення ваг при заданому параметрі швидкості навчання η, що може приймати значення, наприклад, $\eta = 0.1$:

$$w_i^{new} = w_i - \eta \cdot \frac{dL}{dw_i}, i = \overline{1,4}$$

$$b^{new} = b - \eta \cdot \frac{dL}{db},$$

де $\omega_i{}^{new}$ b^{new} – оновлені значення вагових коефіцієнтів на початкового зсуву.

Крок 5. Процес навчання у вигляді повторення кроків 1…4 повторюється щодо множини вхідних векторів поки не буде отримане задане значення помилки.

Перевіримо роботу алгоритму в режимі навчання.

Результати роботи нейромережі в режимі навчання при початкових параметрах (швидкість навчання η=0.1, вагові коефіцієнти $\omega_i = 0.1$, $b = 0.5$) наведені в таблиці 3.

Таблиця 3.

Процес навчання нейромережі

E_m	T_m	R_z	E_z	y	y_a	**Помилка «є» чи «ні»**
0	0	0	0	0.06	0	ні
0	0	0	1	0.03	0	ні
1	1	1	0	0.15	0	ні
1	1	1	1	0.09	0	ні
2	2	2	0	0.33	0	ні
2	2	2	1	0.21	0	ні
3	3	3	0	0.59	0	Є
3	3	4	0	0.44	0	ні
4	3	4	1	0.49	0	ні
4	4	4	0	0.53	1	ні
4	4	4	1	0.80	1	ні
5	5	5	0	0.69	1	ні
5	5	5	1	0.91	1	ні
6	6	6	0	0.86	1	ні
6	6	6	1	0.96	1	ні
7	7	7	0	0.94	1	ні
7	7	7	1	0.98	1	ні
8	8	8	1	0.98	1	ні
9	9	9	1	0.99	1	ні

Графік залежності виходу нейрона від кількості навчальних векторів показаний на рисунку 3.

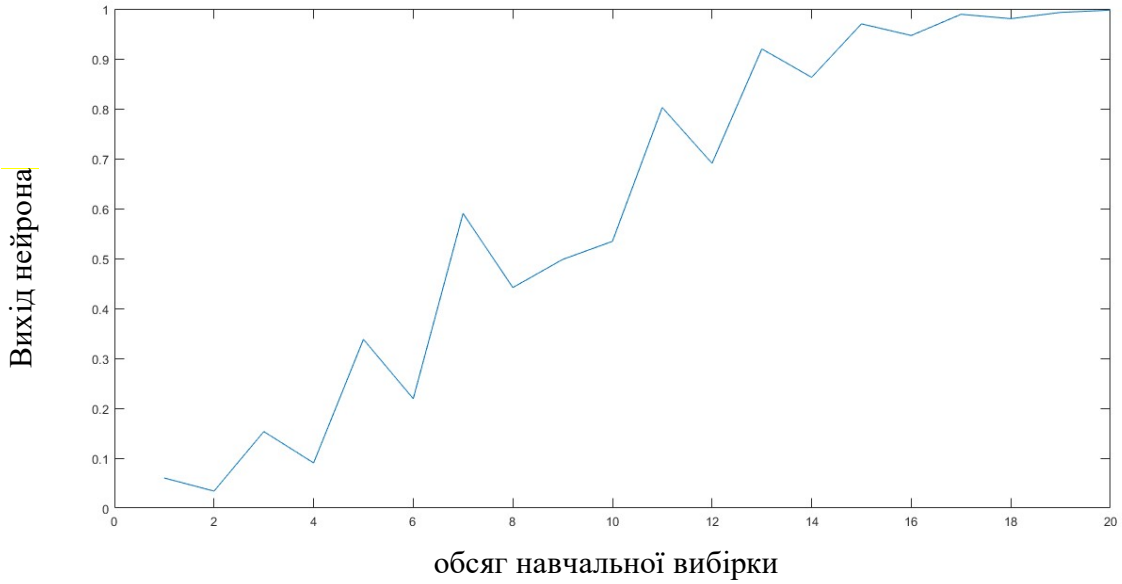

Рис. 3. Залежність виходу нейрона від кількості навчальних векторів

Як можна побачити, помилки ідентифікації спам-контенту відбуваються тільки на ранніх етапах навчання, тобто при недостатньої кількості вхідних навчальних векторів. Можна зробити висновок, що при кількості навчальних векторів більше 10, імовірність помилки різко зменшується і даний алгоритм можна використовувати.

Розглянемо ще одну модифікацію алгоритму виявлення спаму із навчанням за мінімальною довжиною навчальної вибірки.

Алгоритм виявлення фішингу із використанням нейрону із пороговою функцією активації.

Крок 1. На основі аналізу тексту повідомлення формуємо вектор вхідних чинників

$$V_p = \left[R_m, T_m, E_z, R_z\right]$$

де R_m, T_m, E_z можуть приймати дискретні значення з діапазону від 0 до 10,

R_z – 0 чи 10 в залежності від наявності чи відсутності посилань на зовнішні джерела.

Крок 2. Задаймо всі вагові коефіцієнті виконавчого нейрона однаковими $\omega_1 = \omega_2 = \omega_3 = \omega_4 = 10$ і формуємо ваговий вектор, тобто $W = \left[10,10,10,10\right]$.

Крок 3. Обчислюємо значення аргументу p функції активації нейрона:

$$p = \frac{1}{\left(\sum_{i=1}^{n} w_i^2\right)} V_p^T W = \frac{\sum_{i=1}^{n} v_i w_i}{\sum_{i=1}^{n} w_i^2},$$

де v_i – елементи вектору V_p,

ω_i – елементи вектору w,

n – кількість врахованих чинників, в даному випадку n = 4.

Крок 4. Обчислюємо значення вихідного сигналу виконавчого нейрону, тобто значення функції активації $F_A(p)$ в залежності від значення аргументу p.

У якості функції активації обираємо порогову функцію:

$$F_A(p) = \begin{cases} 1, & p \geq 0.5 \\ 0, & p < 0.5 \end{cases}$$

Якщо отримано значення «1» то текст листа містить фішинг, а якщо «0» – то цей лист етичний.

Крок 5. Якщо був активований режим навчання то переходимо на крок 6, інакше алгоритм завершуємо.

Крок 6. Перевіряємо чи вірно виявлена наявність чи відсутність фішингу? Якщо вірно процес навчання завершуємо, інакше переходимо на крок 7.

Крок 7. Обчислюємо величину похибки:

$$E = \left|0.5 - p\right| = \left|0.5 - \frac{\sum_{i=1}^{4} v_i w_i}{\sum_{i=1}^{4} w_i^2}\right|$$

Крок 8. Розраховуємо нове значення чинника R_z, R_{zc}

Якщо не було виявлено фішинг, тобто хибно прийнято рішення про відсутність фішингу:

$$R_{zc} = R_{zp} + \left|0.5 - p\right| \frac{1}{\max(R_m, T_m, E_z)} \sum_{i=1}^{4} w_i^2 = R_{zp} + E \frac{1}{\max(R_m, T_m, E_z)} \sum_{i=1}^{4} w_i^2,$$

де R_{zp} – попереднє значення чинника R_z;

max (R_m, T_m, E_z) – максимальне значення по всім чинникам.

Якщо було хибно виявлено фішинг, при умові його відсутності:

$$R_{zc} = R_{zp} - \left|0.5 - p\right| \frac{1}{\max(R_m, T_m, E_z)} \sum_{i=1}^{4} w_i^2 = R_{zp} - E \frac{1}{\max(R_m, T_m, E_z)} \sum_{i=1}^{4} w_i^2,$$

Крок 9. Формуємо новий вектор вхідних даних $V_p = [R_m, T_m, E_z, R_{zc}]$ і переходимо на крок 2.

Перевіримо роботу алгоритму.

Хай наявний лист володіє наступними чинниками: $R_m = 5$, $T_m = 7$, $E_z = 3$, $R_z = 10$.

Крок 1. Формуємо вектор вхідних чинників:

$$V_p = [5, 7, 3, 10].$$

Крок 2. Задаймо всі вагові коефіцієнті нейрона однаковими $\omega_1 = \omega_2 = \omega_3 = \omega_4 = 10$ і формуємо ваговий вектор, тобто $W = [10, 10, 10, 10]$.

Крок 3. Обчислюємо значення аргументу p функції активації нейрону:

$$p = \frac{1}{\left(\sum_{i=1}^{4} w_i^2\right)} V_p^T W = \sum_{i=1}^{4} v_i w_i = \frac{50 + 70 + 30 + 100}{400} = 0.625$$

Крок 4. Обчислюємо значення функції активації: $F_A(0.625) = 1$.

Робимо вірний висновок про те, що текст повідомлення є спам-фішингом.

Сформуємо матрицю S_p взаємних впливів вхідних чинників листу, що аналізується за наступним правилом:

$$S_p = V_p^T * V_p = \begin{bmatrix} R_m \\ T_m \\ E_z \\ R_z \end{bmatrix} * [R_m, T_m, E_z, R_z] = \begin{bmatrix} R_m^2 & R_m T_m & R_m E_z & R_m R_z \\ T_m R_m & T_m^2 & T_m E_z & T_m R_z \\ E_z R_m & E_z T_m & E_z^2 & E_z R_z \\ R_z R_m & R_z T_m & R_z E_z & R_z^2 \end{bmatrix}$$

Елементи матриці S_p – s_{ij} мають фізичний зміст взаємного впливу чинників *i* та *j* один на одного.

Побудуємо матрицю S_p для вищерозглянутого прикладу:

$$S_p = \begin{bmatrix} 5 \\ 7 \\ 3 \\ 10 \end{bmatrix} * [5, 7, 3, 10] = \begin{bmatrix} 25 & 35 & 15 & 50 \\ 35 & 49 & 21 & 70 \\ 15 & 21 & 9 & 30 \\ 50 & 70 & 30 & 100 \end{bmatrix}$$

Максимальне значення елементів, що не знаходіться на головної діагоналі $s_{24} = s_{42}$, тобто чинники часу та ризику мають найбільший вплив один на одного.

Розглянемо тепер етичний лист, текст якого має наступні чинники:

$R_m = 1$– «оплата виконаної роботи гарантована»;

$T_m = 1$ – «…до визначеної дати»;

$E_z = 4$ – «Вам надана можливість отримати грантову підтримку».

Сформуємо вектор чинників V_e вважаючи $R_z = 0$, тобто $V_e = [6, 2, 4, 0]$.

Знайдемо значення аргументу функції активації нейрону:

$$p = \frac{60 + 20 + 40}{400} = 0.3$$

Значення функції активації: F(p)= F(0.3)=0. Таким чином спам чи фішинг вірно не виявлено, лист є етичним.

Максимальне значення елементів, що не знаходіться на головної діагоналі $s_{13} = s_{31}$, тобто чинники винагороди чи плати за виконану роботу та емоційного забарвлення мають найбільший вплив один на одного в більшості етичних листів.

Побудуємо матрицю S_p для чинників етичного листу:

$$S_p = \begin{bmatrix} 6 \\ 2 \\ 4 \\ 0 \end{bmatrix} * \begin{bmatrix} 6 & 2 & 4 & 0 \end{bmatrix} = \begin{bmatrix} 36 & 12 & 24 & 0 \\ 12 & 4 & 8 & 0 \\ 24 & 8 & 16 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Висновки. В даної роботі було проаналізовано найбільш агресивні види спаму – фішингу, що можуть призвести до значних фінансових втрат. Було виявлено, що спам-листи та етичні листи схожої спрямованості, наприклад, грантові пропозиції чи виграши в лотерею мають спільні чинники, але їхнє співвідношення та вплив на кінцевий результат відрізняються. Було виділено такі загальні чинники як «можлива винагорода», «обмеження у часі», «емоційне забарвлення» та «наявність посилання на зовнішній ресурс». Щодо перших 3-х чинників запропонована 10-ти бальна шкала оцінки впливу згідно аналізу наявності шаблонів-словосполучень, а «наявність посилання на зовнішній ресурс» має дві градації 1, якщо посилання є та 0 якщо його немає. Сформований вектор чинників, що можуть вказувати на наявність фішингу був оброблений за допомогою двох алгоритмів на базі одношарової нейромережі – із класичною сигмоїдною функцією активації та із пороговою функцією активації. Нейромережа із сигмоїдною функцією активації показує більшу точність та дозволяє задавати діапазон невизначеності, що можна класифікувати як «підозра на фішинг», однак потребує суттєвого збільшення обсягу навчаючої вибірки. Мінімальний обсяг навчальної вибірки для забезпечення імовірності помилки не менш 10^{-3} – по десять листів фішингових та етичних. Проте нейромережа із пороговою функцію значне простіша в реалізації. Запропоновані алгоритми можна удосконалювати при збільшенні чинників впливу, що підлягають токенізації та ранжуванню.

Список літератури

1. G. V. Cormack, *Email spam filtering: A systematic review*, vol. 1, no. 4. 2006.
2. E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa,“Machine learning for email spam filtering: review, approaches and open research problems,” *Heliyon*, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.
3. D. Gaurav, S. M. Tiwari, A. Goyal, N. Gandhi, and A. Abraham, “Machine intelligence-based algorithms for spam filtering on document labeling,” *Soft Comput*., vol. 24, no. 13, pp. 9625–9638, 2020, doi: 10.1007/s00500-019-04473-7.
4. N. Kumar, S. Sonowal, and Nishant, “Email Spam Detection Using Machine Learning Algorithms,”*Proc. 2nd Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2020*, no. July 2020, pp. 108–113, 2020, doi:10.1109/ICIRCA48905.2020.9183098.
5. I. AbdulNabi and Q. Yaseen, “Spam email detection using deep learning techniques,” *Procedia Comput. Sci.,* vol. 184, no. 2019, pp. 853–858, 2021, doi: 10.1016/j.procs.2021.03.107.
6. S. M. Yasin and I. H. Azmi, “Email Spam Filtering Technique: Challenges and Solutions,” *J. Theor. Appl. Inf. Technol.,* vol. 101, no. 13, pp. 5130–5138, 2023
7. A. S. Rajput, V. Athavale, and S. Mittal, “Intelligent model for classification of SPAM and HAM,” *Int. J. Innov. Technol. Explor. Eng*., vol. 8, no. 6, pp. 773–777, 2019
8. G. Waja, G. Patil, C. Mehta, and S. Patil, “How AI Can be Used for Governance of Messaging Services: A Study on Spam Classification Leveraging Multi-Channel Convolutional Neural Network,” *Int. J. Inf. Manag. Data Insights*, vol. 3, no. 1, p. 100147, 2023, doi:10.1016/j.jjimei.2022.100147
9. Blacklisting vs. Whitelisting. — https://consoltech.com/blog/blacklisting-vs-whitelisting/.
10. REALTIME BLACKHOLE LIST (RBL): WHAT IS IT? https://www.mailinblack.com/en/ressources/blog/rbl-what-isit/.
11. N. N. Amir Sjarif, N. F. Mohd Azmi, S. Chuprat, H. M. Sarkan, Y. Yahya, and S. M. Sam, “SMS spam message detection using term frequency-inverse document frequency and random forest algorithm,” *Procedia Comput. Sci*., vol. 161, pp. 509–515, 2019, doi:

10.1016/j.procs.2019.11.150.
12. Wu J, Deng T. Research in Anti-Spam Method Based on Bayesian Filtering. IEEE, Pacific-Asia Workshop on Computational Intelligence and Industrial Application, pp. 887 – 891, 2008
13. Ozgur L, Gungor T, Gurgen F. Spam Mail Detection Using Artificial Neural Network and Bayesian Filter. Springer, pp. 505-510, 2004.
14. Dong J, Cao H, Liu P, Ren L (2006). Bayesian Chinese Spam Filter Based on Crossed N-gram. Intelligent Systems Design and Applications, 2006. ISDA '06. Sixth International Conference on, 3:103-108.
15. Naive Bayes spam filtering. https://en.wikipedia.org/wiki/Naive_Bayes_ spam_filtering.
16. An Approach to Email Classification Using Bayesian Theorem. https://core.ac.uk/download/pdf/231152837.pdf.
17. Ten Spam-Filtering Methods Explained.: https://www.connectingup.org/learn/articles/ten-spam-filtering-methodsexplained.
18. Spam Filtering Using Bag-of-Words. https://heartbeat.fritz.ai/spam-filtering-using-bag-of-words-1c5484ff07f1.
19. Mohammed MA, Gunasekaran SS, Mostafa SA, Mustafa A, Ghani MKA. Implementing an agent-based multi-natural language anti-spam model. 2018 International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR). Putrajaya, Malaysia: IEEE; 2018 Aug. p. 1–5.
20. Akinyelu AA, Adewumi AO. Classification of phishing email using random forest machine learning technique. J Appl Math.2014;2014:425731.
21. Sutta N, Liu Z, Zhang X. A study of machine learning algorithms on email spam classification. *Proceedings of 35th International Conference*. Vol. 69. Missouri, USA: Southeast Missouri State University; 2020. p. 170–9
22. Khamis SA, Foozy CFM, Ab Aziz MF, Rahim N. Header based email spam detection framework using support vector machine (SVM) technique. *International Conference on Soft Computing and Data Mining*. Cham: Springer; 2020 Jan. p. 57–65.
23. S. Chaudhry, S. Dhawan, R. Tanwar, Spam Detection in Social Network Using Machine Learning Approach, in: U. Batra, N. Roy, B. Panda (Eds.), Data Science and Analytics. REDSET 2019, Communications in Computer and Information Science, 2020, pp. 236-245.doi:10.1007/978-981-15-5830-6_20.
24. Ch. Zhao, Y. Xin, X. Li, Y. Yang, Yu. Chen, A Heterogeneous Ensemble Learning Framework for Spam Detection in Social Networks with Imbalanced Data, Applied Sciences, 10, 936, 2020. doi:10.3390/app10030936.
25. N. Sharma, Understanding the Mathematics behind Support Vector Machines, Heartbeat, 2020.URL: https://heartbeat.fritz.ai/understanding-the-mathematics-behind-support-vector-machines-5e20243d64d5.
26. Converting Raw Text to Numerical Vectors using Bag of Words, N Grams and TF-IDF. https://aiaspirant.com/bag-ofwords/.
27. Метод опорних векторів. Режим доступу: https://uk.wikipedia.org/wiki/ Метод_опорних_векторів.

PHISHING DETECTION ALGORITHM IN MESSENGER LETTERS AND EMAILS USING NEURAL NETWORKS WITH A FIXED NUMBER OF RANKING RISK FACTORS

V.O. Nazarov, A.V. Sadchenko, O.A. Kushnirenko

National Odesa Polytechnic University
1, Shevchenko Ave., Odesa, 65044, Ukraine
Emails; koa@op.edu.ua, sadchenko@op.edu.ua

The article is devoted to the development of algorithms for detecting the most harmful type of spam - phishing - based on neural networks with sigmoid and threshold activation functions. A review and analysis of existing approaches to spam detection was conducted and it was found that most modern algorithms use a quantitative approach based on the constant replenishment of dictionaries of word combinations in explicit and tokenized form, which, when using, for example, a neural network at the decision-making stage, can increase the probability of classifying ethical content of messages and letters as spam. A new approach to forming a vector of input factors that can be present in both spam and ethical content is presented. Four main general factors were identified, ranked on a ten-point risk scale: “possible reward”, “time limit”, “emotional coloring” and “presence of a link to an external resource”. Regarding the first 3 factors, a 10-point scale of impact assessment is proposed according to the analysis of the presence of word-combination patterns, and “presence of a link to an external resource” has two gradations: “1” if there is a link, and “0” if there is no link. This ranking allowed us to create a digital representation of risk factors in the form of an input vector of factors, which is then fed to the input of a neural network with sigmoid or threshold activation functions. A neural network using a sigmoid activation function after training, based on ten cases of phishing content and the same number of ethical emails, provided a probability of detecting malicious content of approximately 10^{-3}, provided that suspicious emails are classified as spam. A neural network with a threshold activation function and an additional division of the range of the factor “presence of a link to an external resource” into ten conditional gradations, using the analysis of the S matrix of mutual influence, showed a high learning rate and a probability of detecting malicious content of approximately $0.5 \cdot 10^{-3}$. The results of the research can be used to create phishing filters with both hard and soft solutions, which are intended for integration into email services and social network messengers.

Keywords: spam filters, phishing emails, neural network with sigmoid activation function, risk factors.