eoss-conf.com

ISSUE
№97

EUROPEAN OPEN
SCIENCE SPACE

COLLECTION OF SCIENTIFIC PAPERS

5TH INTERNATIONAL
SCIENTIFIC
AND PRACTICAL
CONFERENCE

SCIENTIFIC RESEARCH: MODERN INNOVATIONS AND FUTURE PERSPECTIVES

JULY 20-22, 2026, MONTREAL, CANADA

EUROPEAN OPEN SCIENCE SPACE

Proceedings of the 5th International Scientific and Practical Conference

"Scientific Research: Modern Innovations and Future Perspectives"

July 20-22, 2026

Montreal, Canada

Collection of Scientific Papers

Canada, 2026

UDC 01.1

Collection of Scientific Papers with the Proceedings of the 5th International Scientific and Practical Conference «Scientific Research: Modern Innovations and Future Perspectives» (July 20-22, 2026, Montreal, Canada). European Open Science Space. 2026.

ISBN 979-8-89704-956-1 (series)
DOI 10.70286/EOSS-20.07.2026

The conference is included in the Academic Research Index ReserchBib International catalog of scientific conferences.

The conference is registered in the database of scientific and technical events of UkrISTEI to be held on the territory of Ukraine (Certificate №1074 dated 22.12.2025).

The materials of the collection are presented in the author's edition and printed in the original language. The authors of the published materials bear full responsibility for the authenticity of the given facts, proper names, geographical names, quotations, economic and statistical data, industry terminology, and other information.

ISBN 979-8-89704-956-1 (series)

EUROPEAN OPEN SCIENCE SPACE

Section: History and Cultural Studies

Section: Information Technology, Cyber Security and Computer Engineering

DOI 10.70286/EOSS-20.07.2026.006.109-128

ЕВОЛЮЦІЯ ТЕКСТОВИХ МОДЕЛЕЙ: РЕКУРЕНТНІ НЕЙРОННІ МЕРЕЖІ В КОНТЕКСТІ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ

Почанський Олег
к.т.н.
Кафедра кібербезпеки та інформаційних технологій
Харківський національний економічний університет
імені С. Кузнеця, Україна

Анотація

У статті проведено комплексний аналіз методів автоматичної генерації тексту за допомогою рекурентних нейронних мереж (RNN) та моделей довгої короткочасної пам'яті (LSTM), а також досліджено сучасні інтелектуальні й інфраструктурні технології протидії спаму. Розглянуто структурні особливості LSTM-мереж та їхню здатність моделювати довгострокові контекстні залежності при створенні текстів. Систематизовано актуальні категорії спаму та здійснено порівняльний аналіз технічних засобів захисту електронної пошти, зокрема криптографічних протоколів автентифікації відправників SPF, DKIM, DMARC та методу сірого списку (greylisting). Обґрунтовано математичний підхід до застосування наївного байєсівського класифікатора для контентної фільтрації повідомлень. Результати роботи мають практичне значення для проектування комбінованих систем кібербезпеки та захисту інформаційних комунікацій від автоматизованого спаму і фішингу.

Ключові слова: автоматична генерація тексту, рекурентні нейронні мережі (RNN), LSTM, класифікація спаму, інформаційна безпека, SPF, DKIM, DMARC, наївний байєсівський класифікатор.

1 Вступ

Обґрунтування актуальності теми. Сучасний етап розвитку глобального інформаційного простору характеризується стрімким збільшенням обсягів текстових даних, що циркулюють у системах електронних комунікацій. Автоматизація процесів обробки природної мови (NLP) та вдосконалення моделей штучного інтелекту відкрили нові можливості для автоматичної генерації текстів. Проте, поряд із корисними сферами застосування, технології автоматичного синтезу тексту все частіше використовуються зловмисниками для створення високореалістичного, унікального та масового небажаного контенту (спаму), фішингових повідомлень та засобів соціальної інженерії.

Електронна пошта залишається одним із ключових каналів ділової та особистої комунікації, але водночас є головним вектором кібератак. Традиційні методи фільтрації, що базуються на простих сигнатурах або статичних правилах,

виявляються неефективними проти динамічно генерованого спаму. Це зумовлює гостру потребу в детальному дослідженні як самих методів автоматичної генерації тексту (зокрема, на основі нейромережевих архітектур), так і комплексних, інтелектуальних засобів протидії кіберзагрозам у поштових системах.

Аналіз останніх досліджень і публікацій. Питання моделювання послідовностей даних та аналізу природної мови висвітлені у фундаментальних працях багатьох дослідників у галузі машинного навчання. Значний внесок у розвиток рекурентних нейронних мереж (RNN) та розв'язання проблеми затухання градієнтів за допомогою архітектури довгої короткочасної пам’яті (LSTM) зробили С. Хохрайтер (S. Hochreiter) та Ю. Шмідхубер (J. Schmidhuber). Проблематика верифікації відправників електронної пошти та стандартизації криптографічних протоколів безпеки (SPF, DKIM, DMARC) детально описана у специфікаціях RFC та дослідженнях фахівців з інформаційної безпеки, таких як М. Кучерава (M. Kucherawy) та Т. Зінк (T. Zink). Водночас статистичні підходи до контентної фільтрації на основі Наївного байєсівського класифікатора залишаються класичним та ефективним базисом для сучасних антиспам-систем. Проте швидка еволюція методів штучної генерації тексту вимагає постійного переосмислення та інтеграції цих підходів.

Виділення невирішених раніше частин загальної проблеми. Попри наявність значної кількості досліджень, присвячених окремо нейромережевим моделям мови та окремо засобам поштової безпеки, існує потреба в комплексному аналізі взаємозв'язку між методами автоматичного створення контенту та механізмами його виявлення. Зокрема, недостатньо систематизовано порівняльний аналіз технічних рівнів захисту (інфраструктурних протоколів автентифікації) та інтелектуальних рівнів (контентного аналізу за допомогою статистичних методів) в умовах протидії автоматизованому спаму.

2 Мета та задачі дослідження

Метою роботи є систематизація та комплексне теоретичне дослідження методів автоматичної генерації тексту на основі рекурентних нейронних мереж (RNN та LSTM), аналіз сучасних загрозоутворюючих категорій спаму, а також оцінка ефективності інфраструктурних (SPF, DKIM, DMARC) та інтелектуальних (байєсівська фільтрація) методів захисту інформаційних систем від небажаної кореспонденції.

Для досягнення поставленої мети необхідно вирішити такі завдання:

розглянути математичні та структурні особливості функціонування базових рекурентних нейронних мереж та моделей LSTM;

проаналізувати основні категорії та технічні аспекти поширення сучасного спаму (фішинг, підробка пошти, масові розсилки);

дослідити специфіку роботи та порівняти ефективність криптографічних протоколів автентифікації джерел повідомлень;

обґрунтувати математичний підхід до використання наївного байєсівського класифікатора для контентного аналізу електронних листів.

Методологічну основу дослідження становлять методи штучного інтелекту, теорія нейронних мереж, теорія ймовірностей та математична статистика, а також системний аналіз у сфері кібербезпеки та захисту інформації.

3 Результати дослідження та обговорення

3.1 Рекурентна нейронна мережа

Рекурентна нейронна мережа (RNN) – це тип нейронних мереж із петлями, які дозволяють зберігати інформацію (рис. 1).

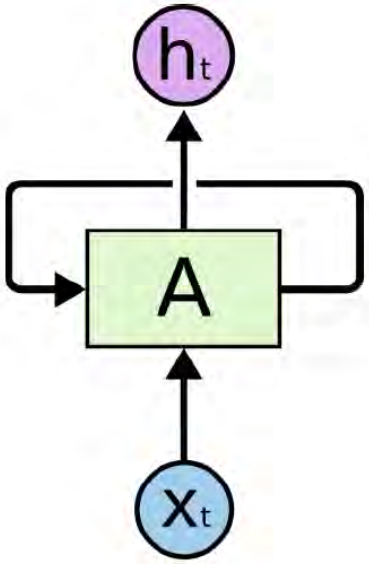


Рисунок 1. Базова структура циклу в рекурентної нейронної мережі [1]

На рис. 1 центральний квадрат представляє нейронну мережу, яка приймає на вхід x_t на поточному зрізі часу t і надає значення h_t як вихід. Цей цикл дає змогу надсилати інформацію від одного кроку мережі до наступного.

Наявність циклічних зв'язків є характерною особливістю рекурентних нейронних мереж і забезпечує можливість врахування інформації з попередніх кроків обробки даних. Незважаючи на складність такої архітектури, рекурентну нейронну мережу можна інтерпретувати як послідовність однакових нейронних блоків, які мають спільні параметри та передають інформацію між сусідніми часовими кроками. У процесі розгортання рекурентної мережі в часі її структура подається у вигляді ланцюжка взаємопов'язаних копій одного й того самого обчислювального модуля, що дозволяє наочно відобразити механізм передачі інформації та залежностей між послідовними елементами вхідних даних (рис. 2).

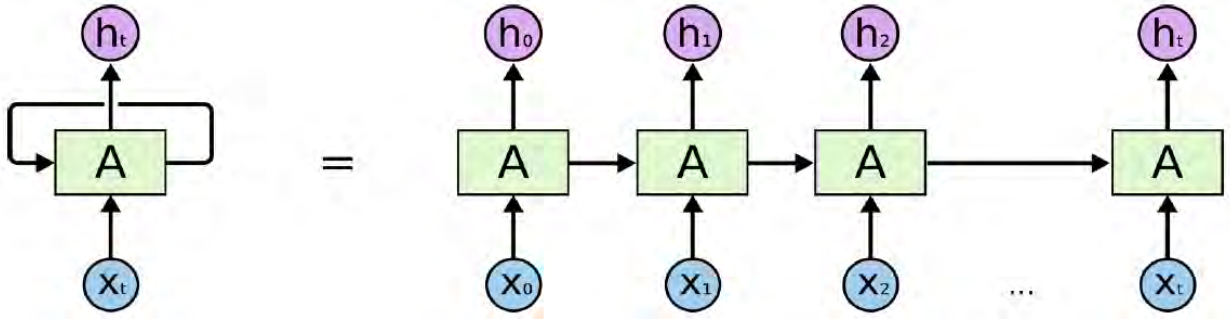


Рисунок 2. Розгорнута рекурентна нейронна мережа [1]

Розгорнута структура рекурентної нейронної мережі демонструє її природну орієнтованість на обробку послідовних даних. Завдяки механізму передачі інформації між часовими кроками такі мережі здатні враховувати контекст попередніх елементів послідовності, що робить їх ефективним інструментом для аналізу текстів, часових рядів та інших впорядкованих наборів даних.

Протягом останніх років рекурентні нейронні мережі успішно застосовуються для розв'язання широкого спектра завдань, серед яких розпізнавання мовлення, моделювання природної мови, машинний переклад, аналіз текстової інформації та прогнозування послідовностей. Значний внесок у підвищення ефективності цих систем зробила поява архітектури Long Short-Term Memory (LSTM), яка є спеціалізованим різновидом рекурентних нейронних мереж.

На відміну від класичних RNN, мережі LSTM здатні ефективніше зберігати та використовувати інформацію про довгострокові залежності в даних завдяки наявності механізму керування пам'яттю. Це дозволяє уникати проблеми зникнення градієнта та суттєво покращує якість навчання на довгих послідовностях. Саме тому більшість сучасних рішень, що базуються на рекурентних нейронних мережах і демонструють високі результати в задачах обробки природної мови, використовують архітектуру LSTM як основний інструмент моделювання послідовних даних.

3.2 LSTM мережа

Мережі довгої короткочасної пам'яті (Long Short-Term Memory, LSTM) є спеціалізованим різновидом рекурентних нейронних мереж, призначеним для моделювання та аналізу довгострокових залежностей у послідовних даних. Завдяки своїй архітектурі вони здатні ефективно зберігати та використовувати інформацію протягом тривалих часових інтервалів, що робить їх придатними для розв'язання широкого спектра завдань, пов'язаних з обробкою природної мови, розпізнаванням мовлення, машинним перекладом та аналізом текстової інформації.

Архітектура LSTM була розроблена для подолання обмежень класичних рекурентних нейронних мереж, зокрема проблеми зникнення та вибуху градієнтів під час навчання. На відміну від традиційних RNN, мережі LSTM містять спеціальні механізми пам'яті та керування потоками інформації, що дозволяє зберігати важливі дані протягом тривалого часу та відкидати несуттєву інформацію.

Як і всі рекурентні нейронні мережі, LSTM можуть бути представлені у вигляді послідовності взаємопов'язаних повторюваних модулів (рис. 3). Однак якщо у класичних RNN кожен модуль зазвичай складається з одного нейронного шару з функцією активації, наприклад гіперболічним тангенсом, то в архітектурі LSTM повторюваний модуль має значно складнішу структуру. Він містить комірку пам'яті та набір керуючих механізмів (вхідні, вихідні та забувальні ворота), які забезпечують ефективне збереження, оновлення та передачу інформації між часовими кроками.

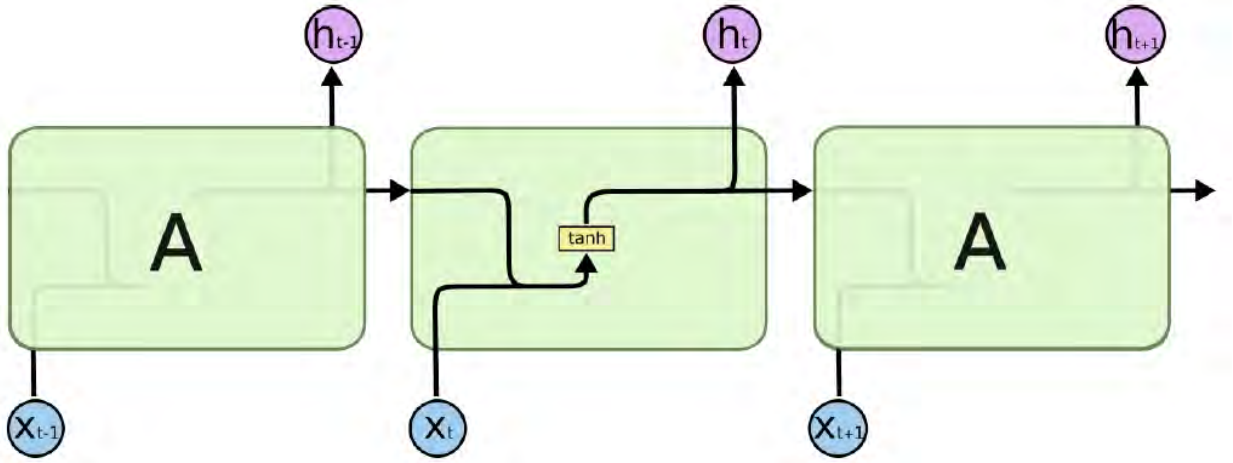


Рисунок 3. Повторюваний модуль у стандартній RNN містить один шар [1]

Мережі LSTM, подібно до інших рекурентних нейронних мереж, мають послідовну ланцюгову структуру, у якій інформація передається між послідовними часовими кроками. Проте на відміну від класичних рекурентних мереж, де повторюваний модуль зазвичай представлений одним нейронним шаром, архітектура LSTM характеризується значно складнішою внутрішньою будовою (рис. 4). Кожен повторюваний модуль LSTM складається з декількох взаємопов'язаних обчислювальних компонентів, які спільно забезпечують керування процесами збереження, оновлення та передавання інформації. Зокрема, структура модуля включає комірку пам'яті та систему спеціальних керуючих воріт, що дозволяють вибірково зберігати важливі дані, відкидати неактуальну інформацію та формувати вихідний сигнал. Така організація забезпечує ефективне моделювання довгострокових залежностей і підвищує якість обробки послідовних даних порівняно з традиційними рекурентними нейронними мережами.

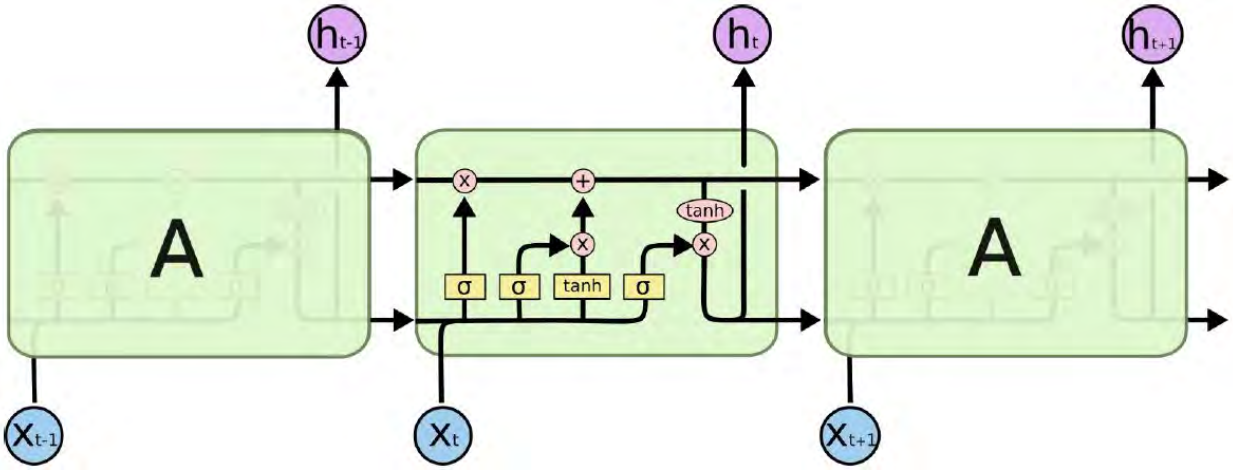


Рисунок 4. LSTM з чотирма взаємодіючими шарами [1]

На рис. 4 представлено архітектуру мережі LSTM у вигляді послідовності взаємопов'язаних модулів, між якими передаються вектори даних. Інформаційні потоки відображаються лініями, що з'єднують виходи одного модуля з входами наступних. На схемі окремі обчислювальні операції позначені круглими

елементами, тоді як прямокутні блоки відповідають шарам нейронної мережі. Операції об'єднання даних відображаються точками злиття потоків, а розгалуження ліній вказують на копіювання та передачу інформації до кількох компонентів моделі одночасно.

Центральним елементом архітектури LSTM є стан комірки пам'яті (cell state), який проходить через усю мережу вздовж її часової осі. Стан комірки виконує роль каналу довготривалого зберігання інформації та забезпечує передавання важливих даних між послідовними часовими кроками (рис. 5). Завдяки переважно лінійному характеру поширення інформації через цей канал втрати корисних даних мінімізуються, що дозволяє мережі ефективно зберігати відомості про попередні події протягом тривалих інтервалів часу. Саме ця особливість забезпечує здатність LSTM моделювати довгострокові залежності в послідовних даних та підвищує ефективність навчання порівняно з класичними рекурентними нейронними мережами.

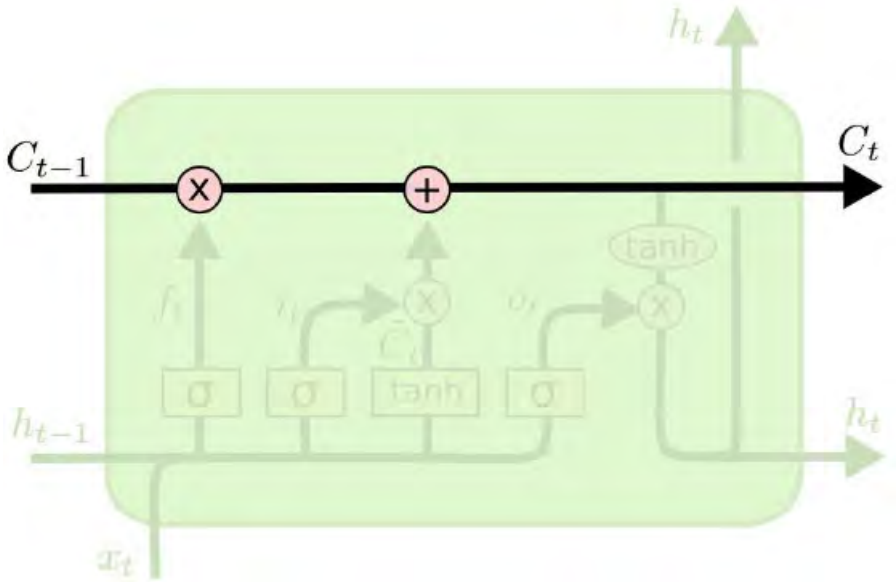


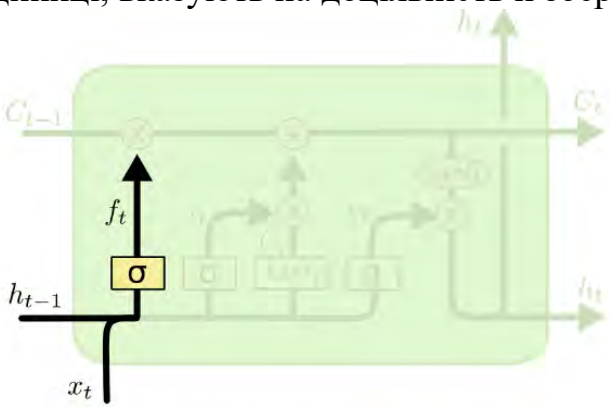
Рисунок 5. Стан комірки [1]

Однією з ключових переваг архітектури LSTM є наявність механізму керування інформацією, що зберігається в комірці пам'яті. На відміну від класичних рекурентних нейронних мереж, LSTM здатна не лише накопичувати інформацію, а й вибірково оновлювати або видаляти її залежно від поточного контексту. Цей процес реалізується за допомогою спеціальних структур, які називаються воротами (gates).

Ворота виконують функцію контролю потоків інформації всередині мережі. Вони складаються із сигмоїдного шару та операцій поелементного множення. Сигмоїдна функція активації формує значення в діапазоні від 0 до 1 для кожного елемента вхідного вектора. Отримані значення визначають ступінь пропускання інформації через відповідний вузол: значення, близькі до 0, практично блокують передачу даних, тоді як значення, близькі до 1, забезпечують їх повне збереження та подальше використання.

Архітектура LSTM включає три основні типи воріт: ворота забування (forget gate), вхідні ворота (input gate) та вихідні ворота (output gate). Їх спільна робота забезпечує ефективне керування станом комірки пам'яті та дозволяє мережі адаптивно реагувати на зміни в послідовності даних.

Першим етапом обробки інформації є визначення того, які дані зі стану комірки пам'яті необхідно зберегти, а які слід видалити (рис. 6). Для цього використовується механізм воріт забування. Ворота забування аналізують прихований стан попереднього кроку (h_{t-1}) та поточний вхідний вектор (x_t), після чого формують вектор коефіцієнтів у діапазоні від 0 до 1 для кожного елемента попереднього стану комірки (C_{t-1}). Значення, близькі до нуля, означають необхідність видалення відповідної інформації, тоді як значення, близькі до одиниці, вказують на доцільність її збереження для подальшої обробки.



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

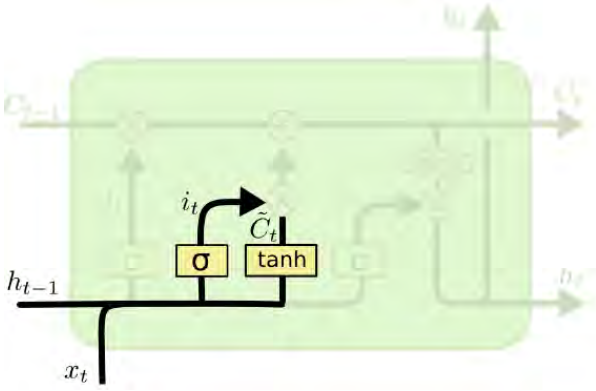
Рисунок 6. Крок перший LSTM моделі [1]

Після виконання операції забування наступним етапом є визначення інформації, яка має бути додана до стану комірки пам'яті. Цей процес реалізується у два послідовні кроки.

Спочатку вхідні ворота (input gate), що реалізуються за допомогою сигмоїдного шару, аналізують поточний вхідний вектор (x_t) та прихований стан попереднього кроку (h_{t-1}). На основі цього аналізу формується вектор коефіцієнтів, який визначає, які компоненти стану комірки підлягають оновленню та з якою інтенсивністю.

Паралельно з роботою вхідних воріт шар з функцією активації гіперболічного тангенса генерує вектор нових кандидатних значень ($\tilde{C}_t$), що містить потенційну інформацію для додавання до комірки пам'яті. Ці значення відображають нові знання, отримані з поточних вхідних даних, та можуть бути використані для оновлення внутрішнього стану мережі.

На наступному етапі результати роботи вхідних воріт та вектор кандидатних значень об'єднуються, що дозволяє сформувати оновлений стан комірки пам'яті. Такий механізм забезпечує контрольоване внесення нової інформації та підтримує здатність мережі ефективно накопичувати знання про довгострокові залежності в послідовних даних (рис. 7).



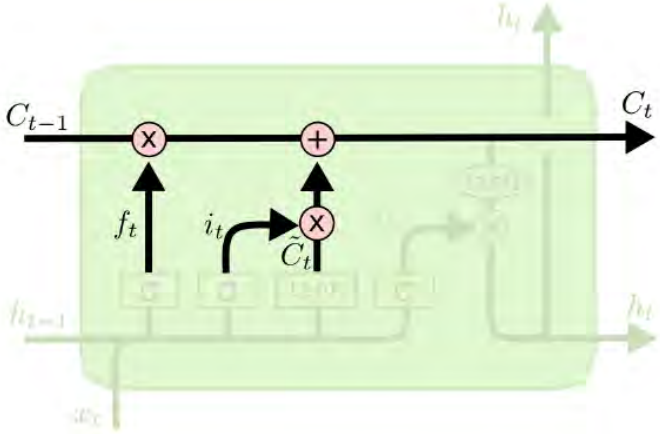
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Рисунок 7. Крок другий LSTM моделі [1]

Після визначення інформації, яку необхідно зберегти, видалити або додати, виконується оновлення стану комірки пам'яті. На цьому етапі попередній стан комірки (C_{t-1}) перетворюється на новий стан (C_t) відповідно до рішень, прийнятих механізмами воріт забування та вхідних воріт.

Спочатку попередній стан комірки множиться на вектор виходу воріт забування (f_t). У результаті цього кроку зберігається лише та інформація, яка була визнана важливою для подальшої обробки, тоді як несуттєві або застарілі дані поступово усуваються. Далі до отриманого результату додається добуток вектора вхідних воріт (i_t) та вектора кандидатних значень ($\tilde{C}_t$). Таким чином, нова інформація включається до стану комірки лише в тих компонентах, для яких вхідні ворота дозволили оновлення.

Завдяки такому механізму мережа LSTM здатна одночасно підтримувати важливу історичну інформацію та інтегрувати нові знання, що забезпечує ефективне моделювання довгострокових залежностей у послідовних даних (рис. 8).



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Рисунок 8. Третій крок LSTM моделі [1]

3.3 Теоретичні основи спаму та його класифікація

Спамом називають масове розповсюдження небажаних електронних повідомлень за допомогою засобів електронного зв'язку без попередньої згоди або запиту з боку одержувачів. Найчастіше спам поширюється через електронну пошту, однак для його розсилання можуть використовуватися й інші

комунікаційні платформи, зокрема месенджери, соціальні мережі та SMS-повідомлення. Основною метою спаму є донесення певної інформації до максимально широкої аудиторії незалежно від зацікавленості користувачів у її отриманні.

У сфері електронної пошти поняття спаму зазвичай пов'язують із двома категоріями повідомлень: небажаною комерційною електронною поштою (Unsolicited Commercial E-mail, UCE) та небажаною масовою електронною поштою (Unsolicited Bulk E-mail, UBE). Такі повідомлення можуть містити рекламні матеріали, пропозиції сумнівних послуг, фішингові посилання, шкідливі вкладення або інші елементи, що становлять загрозу для інформаційної безпеки користувачів.

Електронні повідомлення, які належать до категорії спаму, зазвичай характеризуються такими ознаками [2]:

1. Відправляються без попереднього запиту або згоди одержувача.
2. Мають рекламний, комерційний або інший нав'язливий характер.
3. Розповсюджуються одночасно великій кількості адресатів із використанням автоматизованих засобів розсилання.

Наявність зазначених характеристик дозволяє розглядати спам як одну з актуальних проблем сучасних інформаційно-комунікаційних систем, що негативно впливає на ефективність роботи поштових сервісів, збільшує навантаження на мережеву інфраструктуру та створює додаткові ризики для інформаційної безпеки.

Спам становить серйозну проблему для сучасних систем електронної пошти, оскільки призводить до неефективного використання обчислювальних та мережевих ресурсів. Масове надходження небажаних повідомлень збільшує навантаження на поштові сервери, займає дисковий простір для зберігання кореспонденції та ускладнює обробку легітимних повідомлень. Крім того, користувачі змушені витрачати додатковий час на перегляд, сортування та видалення небажаних листів, що негативно впливає на продуктивність їхньої роботи.

Переважна більшість спам-повідомлень орієнтована на конкретних користувачів і розповсюджується за допомогою автоматизованих систем масової розсилки. Для формування баз адрес одержувачів спамери використовують різноманітні методи збору інформації. Зокрема, адреси електронної пошти можуть отримуватися шляхом автоматизованого сканування вебсайтів, форумів, груп новин та інших відкритих джерел мережі Інтернет. Також поширеними є випадки використання викрадених або незаконно отриманих списків розсилки, а також спеціалізованих програмних засобів для пошуку адрес електронної пошти на вебресурсах.

Унаслідок таких дій кількість небажаних повідомлень постійно зростає, що створює додаткові ризики для інформаційної безпеки користувачів та організацій. Саме тому розроблення ефективних механізмів автоматичного

виявлення та фільтрації спаму є одним із важливих напрямів розвитку сучасних систем захисту електронної пошти.

Спам-повідомлення можуть мати різне призначення та використовувати різноманітні методи впливу на користувачів. Залежно від змісту та способу реалізації кіберзагроз доцільно виділити декілька основних категорій спаму:

1. Небажані рекламні повідомлення. Найпоширенішим видом спаму є масові рекламні розсилки, які використовуються для просування товарів, послуг або інформаційних ресурсів. Такі повідомлення часто містять пропозиції щодо придбання медичних препаратів, фінансових послуг, освітніх курсів, програмного забезпечення або інших комерційних продуктів. Основною метою подібних розсилок є привернення уваги потенційних клієнтів шляхом масового поширення рекламного контенту [11].

2. Фішинговий спам. Одним із найнебезпечніших різновидів спаму є фішингові повідомлення, спрямовані на отримання конфіденційної інформації користувачів [3]. Такі електронні листи зазвичай маскуються під офіційні повідомлення банків, платіжних систем, державних установ або популярних інтернет-сервісів [4]. У тексті повідомлення користувача спонукають перейти за посиланням та ввести облікові дані, реквізити платіжних карток або іншу персональну інформацію. Отримані відомості надалі можуть використовуватися для несанкціонованого доступу до облікових записів або фінансових ресурсів жертви.

3. Шахрайські повідомлення типу «Нігерійський лист». Даний вид спаму належить до категорії соціальної інженерії та базується на маніпулюванні довірою користувачів. У таких повідомленнях зловмисники повідомляють про можливість отримання значної грошової винагороди або участі у фінансовій операції за умови попередньої сплати певної суми коштів. Як правило, після встановлення контакту жертву поступово переконують здійснювати додаткові платежі під різними приводами, що зрештою призводить до фінансових втрат.

4. Підробка електронної пошти (E-mail Spoofing) [5]. Цей метод полягає у фальсифікації адреси відправника з метою створення враження, що повідомлення надійшло від надійного джерела. Зловмисники можуть видавати себе за представників відомих компаній, державних органів або фінансових установ для підвищення рівня довіри користувача до отриманого листа. Відсутність у базовому протоколі SMTP вбудованих механізмів автентифікації сприяє використанню такого способу атак. Для протидії підробці відправника застосовуються технології SPF [6], DKIM [7] та DMARC [8].

5. Шкідливі вкладення та троянські програми. Значна частина спам-повідомлень використовується для поширення шкідливого програмного забезпечення. Такі листи можуть містити заражені вкладення або посилання на ресурси, з яких завантажується шкідливий код. Після відкриття файлу або переходу за посиланням на пристрій користувача можуть бути встановлені троянські програми, шпигунське програмне забезпечення або інші види шкідливих компонентів [3, 9].

6. Комерційні інформаційні розсилки. До цієї категорії належать повідомлення, які надсилаються легальними компаніями та сервісами з рекламною або інформаційною метою. Хоча такі розсилки зазвичай не містять шкідливого вмісту, вони можуть сприйматися користувачами як небажані у випадках автоматичного додавання до списків розсилки під час реєстрації на вебресурсах.

7. Фальшиві антивірусні повідомлення. Цей тип спаму ґрунтується на психологічному впливі та використанні страху користувача перед можливим зараженням комп'ютера. У листах повідомляється про нібито виявлені віруси або критичні загрози безпеці. Користувачеві пропонується встановити спеціальне програмне забезпечення для усунення проблеми, яке насправді може містити шкідливий код або використовуватися для викрадення даних [9].

8. Політичний та ідеологічний спам. Окрему категорію становлять повідомлення політичного, пропагандистського або екстремістського характеру. Вони можуть використовуватися для поширення дезінформації, маніпулювання громадською думкою або збору персональних даних користувачів під виглядом участі в опитуваннях, громадських кампаніях чи благодійних проєктах [9].

Різноманітність видів спаму свідчить про необхідність застосування сучасних автоматизованих методів його виявлення. Особливої актуальності набувають системи фільтрації на основі методів машинного навчання та байєсовських класифікаторів, які дозволяють аналізувати зміст повідомлень і ефективно розпізнавати різні категорії небажаної кореспонденції.

3.4 Методи протидії спаму

Sender Policy Framework (SPF) – це технічний метод для запобігання підробці адреси відправника електронного листа. Він виник як метод блокування спаму. Незважаючи на те, що SPF існує з 2006 року, він став запропонованим стандартом у 2014 році після публікацій RFC 7208 [6] та RFC 7372 [12]. Для цього системний адміністратор домену повинен налаштувати запис SPF. Це список уповноважених хостів, яким дозволено використовувати своє доменне ім'я в системі доменних імен (DNS). Приклад роботи SPF можна побачити на рис. 9:

1 Починається з того що відправник намагається надіслати електронне повідомлення одержувачу.

2 Вхідний поштовий сервер отримує електронне повідомлення та отримує ім'я домену, з якого воно було надіслане.

3 Тепер сервер вхідної пошти використовує цю інформацію для пошуку DNS для перевірки, чи є запис SPF для цього домену. Якщо IP-адреса, що надсилається в електронному листі, відповідає будь-якій із вихідних адрес, включених до запису SPF, електронна пошта аутентифікується та доставляється. Якщо збіг адреси неможливий, автентифікація не вдається.

4 Коли автентифікація електронної пошти або автентифікація не вдалася, сервер вхідної пошти може обробляти електронну пошту, виходячи з конкретних правил цієї електронної системи та домену [14]. Значення буде додано до заголовка повідомлення електронної пошти Received-SPF на основі перевірки

SPF і якщо воно пройде чи ні. Це значення може бути прийнять, відхилити, м'яке відхилення(буде прийнято, але позначено як спам), нейтральне ставлення, відкинути повідомлення, помилка або тимчасова помилка, які вказані та пояснено у RFC 7208 [6]. Це можливі результати пошуку DNS та перевірки записів SPF або запиту SPF. Однак від вхідного сервера електронної пошти залежить, які дії слід вжити за результатами оцінки. Наприклад, набір правил може передбачати блокування всієї електронної пошти, яка не передає SPF, але доставляти електронну пошту, яка передає SPF, у поштові скриньки кінцевих користувачів.

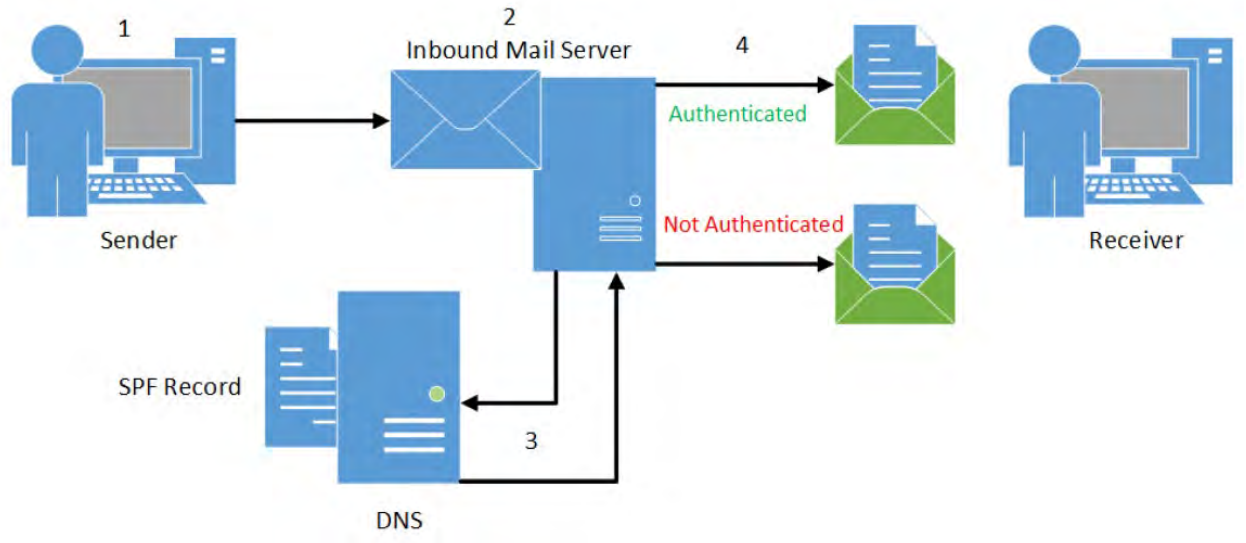


Рисунок 9. Схема роботи SPF [13]

Незважаючи на високу ефективність механізму SPF (Sender Policy Framework) у боротьбі з підробкою адрес відправника, цей підхід має низку обмежень. Однією з проблем є можливість використання зловмисниками власних доменів із коректно налаштованими SPF-записами. У деяких випадках такі домени можуть бути навмисно схожими на домени відомих організацій або сервісів. Оскільки перевірка SPF підтверджує лише право сервера надсилати повідомлення від імені певного домену, а не достовірність самого доменного імені, подібні повідомлення можуть успішно проходити перевірку та потрапляти до папки вхідної пошти користувача.

Таким чином, SPF не є універсальним засобом захисту від спаму та фішингових атак. Його використання дозволяє ускладнити процес підробки електронних адрес і підвищити вартість реалізації таких атак для зловмисників, однак не усуває проблему повністю [15]. Ефективність SPF значною мірою залежить від правильності його налаштування та використання в поєднанні з іншими механізмами автентифікації електронної пошти [16].

Робота SPF безпосередньо пов’язана з функціонуванням системи доменних імен (DNS), оскільки інформація про дозволені сервери відправлення зберігається у вигляді спеціальних DNS-записів. Під час отримання повідомлення поштовий сервер виконує перевірку домену відправника та

аналізує відповідний SPF-запис. У разі недоступності DNS-сервісу або виникнення помилок під час отримання запису процедура перевірки SPF може завершитися невдало або повернути тимчасову помилку, що негативно впливає на процес доставки електронної пошти.

Додатковою проблемою є пересилання електронних повідомлень між поштовими серверами. У таких випадках повідомлення може надсилатися з сервера, який не зазначений у SPF-записі початкового домену, що призводить до помилкового спрацьовування механізму перевірки. Внаслідок цього легітимні повідомлення можуть бути позначені як підозрілі або навіть відхилені приймаючим сервером.

На сьогодні SPF широко використовується як один із базових механізмів автентифікації електронної пошти та захисту від підробки адрес відправника. Його застосування дозволяє організаціям підтверджувати перелік серверів, уповноважених надсилати повідомлення від імені їхніх доменів. У разі виявлення невідповідності між адресою відправника та результатами SPF-перевірки поштові клієнти можуть відображати користувачеві попередження про можливу підробку повідомлення або потенційну шахрайську активність [17].

Водночас надмірно суворе застосування політик SPF на стороні одержувача може спричиняти блокування або помилкове маркування легітимних повідомлень. Саме тому на практиці SPF зазвичай використовується разом із додатковими технологіями захисту, такими як DKIM (DomainKeys Identified Mail) та DMARC (Domain-based Message Authentication, Reporting and Conformance), що забезпечують більш надійну автентифікацію електронної пошти та підвищують ефективність протидії спаму і фішинговим атакам.

Одним із поширених механізмів захисту від підробки електронних повідомлень є технологія DomainKeys Identified Mail (DKIM). DKIM являє собою протокол автентифікації електронної пошти, який використовує криптографію з відкритим ключем для підтвердження достовірності відправника та цілісності повідомлення.

Для функціонування DKIM адміністратор домену повинен опублікувати у системі доменних імен (DNS) спеціальні записи, що містять політику використання DKIM та відкритий ключ домену. Під час надсилання повідомлення поштовий сервер формує цифровий підпис на основі закритого ключа. Після отримання повідомлення сервер одержувача використовує відкритий ключ, опублікований у DNS, для перевірки коректності підпису. У разі успішної перевірки підтверджується, що повідомлення дійсно було надіслано від зазначеного домену та не було змінене під час передавання.

Попри високу ефективність, DKIM має певні недоліки. Зокрема, труднощі виникають під час використання списків розсилки, які можуть змінювати окремі поля заголовка або вміст повідомлення. У такому випадку цифровий підпис перестає відповідати початковому повідомленню, що призводить до помилки автентифікації. Саме складність впровадження та підтримки DKIM є однією з причин його неповного використання окремими організаціями.

Для підвищення надійності автентифікації електронної пошти був розроблений протокол Domain-based Message Authentication, Reporting and Conformance (DMARC). Він поєднує можливості SPF та DKIM і дозволяє визначати політики обробки повідомлень, що не пройшли перевірку автентичності. Крім того, DMARC забезпечує механізм формування звітів, які надсилаються власникам доменів для контролю спроб підробки електронної пошти.

Для успішного проходження перевірки DMARC повідомлення повинно пройти автентифікацію за допомогою SPF або DKIM із додатковою перевіркою відповідності домену відправника, зазначеного в полі «From». Якщо хоча б одна з перевірок завершується успішно та домени узгоджені між собою, повідомлення вважається автентичним. У протилежному випадку до нього можуть бути застосовані політики карантину або відхилення.

Однак DMARC успадковує окремі обмеження SPF та DKIM. Зокрема, можливі випадки помилкового блокування легітимних повідомлень унаслідок некоректної конфігурації системи або особливостей пересилання електронної пошти між серверами.

Поряд із механізмами автентифікації електронної пошти широко використовуються методи фільтрації повідомлень на основі правил.

Одним із найпростіших підходів є використання чорних списків (blacklists). Чорний список являє собою перелік адрес електронної пошти, IP-адрес, доменних імен або інших ознак, пов'язаних із джерелами спаму. Якщо повідомлення містить ознаку, яка входить до чорного списку, воно автоматично позначається як небажане. Чорні списки можуть підтримуватися як окремими користувачами, так і спеціалізованими організаціями, що здійснюють централізований моніторинг джерел спаму.

Протилежним підходом є використання білих списків (whitelists), які містять перелік довірених відправників. Усі повідомлення від адрес, що входять до білого списку, автоматично приймаються системою, тоді як решта можуть бути заблоковані або піддані додатковій перевірці. Недоліком такого підходу є неможливість передбачити всіх потенційних легітимних відправників, що може призводити до втрати корисної кореспонденції.

Ще одним методом є використання сірих списків (greylisting). Принцип роботи полягає в тому, що поштовий сервер тимчасово відхиляє повідомлення від невідомих відправників, повертаючи код тимчасової помилки SMTP. Легітимні поштові сервери зазвичай повторюють спробу доставки через певний проміжок часу, після чого повідомлення приймається. Багато спамерських систем не виконують повторних спроб доставки, тому значна частина спаму відсіюється на цьому етапі.

Найбільш перспективним напрямом автоматичного виявлення спаму є використання статистичних методів машинного навчання. Серед них особливе місце займає наївний байєсовський класифікатор, який завдяки високій швидкодії та точності став одним із найпоширеніших методів фільтрації електронної пошти.

Байєсовський фільтр використовує теорему Байєса для визначення ймовірності того, що повідомлення належить до категорії спаму на основі набору слів, які воно містить. Під час навчання системи формується статистика появи слів у спам-повідомленнях та легітимній кореспонденції. Для цього створюються окремі словники частот для кожного класу повідомлень.

Ймовірність того, що повідомлення є спамом за наявності певного слова, визначається за формулою:

$$P(\text{спам}|\text{слово}) = \frac{P(\text{слово}|\text{спам})P(\text{спам})}{P(\text{слово}|\text{спам})P(\text{спам}) + P(\text{не спам})P(\text{слово}|\text{не спам})}$$

де P(спам|слово) – апостеріорна ймовірність того, що повідомлення є спамом за наявності певного слова;

P(спам) – апріорна ймовірність появи спам-повідомлення;

P(не спам) – апріорна ймовірність появи легітимного повідомлення;

P(слово|спам) – ймовірність появи слова у спам-повідомленні;

P(слово|не спам) – ймовірність появи слова у легітимному повідомленні.

У процесі накопичення статистики система поступово формує модель характерних ознак спаму та звичайної кореспонденції. Завдяки цьому байєсовський класифікатор здатний автоматично адаптуватися до змін у структурі спам-повідомлень та забезпечувати високий рівень точності їх виявлення.

До основних переваг автоматичної фільтрації спаму на основі байєсовського підходу належать такі особливості:

– комплексний аналіз вмісту повідомлення. На відміну від традиційних методів, що базуються на пошуку окремих ключових слів або порівнянні з відомими сигнатурами, байєсовський фільтр аналізує повідомлення в цілому. Під час класифікації враховується сукупність текстових ознак та їх статистичні характеристики, що дозволяє більш точно визначати належність повідомлення до категорії спаму або легітимної кореспонденції.

– зниження кількості хибних спрацьовувань. Байєсовські фільтри навчаються не лише на прикладах спам-повідомлень, але й на легітимних електронних листах, характерних для конкретного користувача або організації. Завдяки цьому система формує індивідуальну модель коректної кореспонденції та значно зменшує ймовірність помилкового віднесення корисних повідомлень до категорії спаму.

– адаптивність до нових видів загроз. Однією з ключових переваг байєсовської фільтрації є здатність до безперервного навчання. У процесі експлуатації система накопичує статистику щодо нових спам-повідомлень і легітимних листів, що дозволяє оперативно адаптувати модель класифікації до змін у характері поштового трафіку та появи нових методів обходу захисних механізмів.

– висока точність класифікації. Використання імовірнісного підходу дає змогу оцінювати внесок кожної ознаки у процес прийняття рішення та визначати

підсумкову ймовірність належності повідомлення до певного класу. Це забезпечує стабільно високі показники точності навіть за умов значних обсягів вхідної кореспонденції.

– можливість персоналізації. Байєсовські фільтри можуть адаптуватися до особливостей листування окремого користувача або організації. У результаті система поступово формує власний набір статистичних закономірностей, що підвищує ефективність фільтрації порівняно з універсальними правилами та сигнатурними методами.

Таким чином, використання байєсовських методів класифікації дозволяє створити гнучку, адаптивну та ефективну систему автоматичного виявлення спаму, здатну забезпечувати високий рівень захисту електронної пошти в умовах постійної еволюції кіберзагроз.

3.5 Аналіз системи фільтрації спаму SpamAssassin

Одним із найбільш відомих і поширених програмних засобів для фільтрації небажаної електронної пошти є програмний комплекс SpamAssassin [19]. Ця система належить до класу правил-орієнтованих спам-фільтрів та використовує механізм оцінювання повідомлень на основі сукупності ознак, кожній з яких призначається певна вага.

Популярність SpamAssassin пояснюється низкою переваг:

1. Широке поширення та перевірена ефективність у реальних поштових системах.
2. Відкритий вихідний код, що забезпечує можливість модифікації та адаптації системи до потреб конкретної організації.
3. Регулярне оновлення баз правил і механізмів виявлення спаму.
4. Підтримка механізмів самонавчання та адаптації до особливостей електронного листування користувачів.

Основу роботи SpamAssassin становить бальна система оцінювання повідомлень. Під час аналізу кожен електронний лист перевіряється на відповідність набору правил, які описують характерні ознаки спаму. Кожне правило має певну вагу, що відображає ступінь його впливу на підсумкове рішення. У разі спрацювання правила до загального рейтингу повідомлення додається або віднімається відповідна кількість балів.

Наприклад, наявність у темі повідомлення типових рекламних слів, написаних великими літерами, може збільшувати підсумковий бал. Водночас ознаки, характерні для легітимної кореспонденції, можуть зменшувати загальну оцінку. Після завершення перевірки отримане значення порівнюється з порогом класифікації. Якщо сумарний бал перевищує встановлений поріг, повідомлення позначається як спам.

Особливістю SpamAssassin є велика кількість критеріїв аналізу. Під час перевірки можуть враховуватися:

- ключові слова та фрази в темі повідомлення;
- лексичні особливості тексту листа;
- структура HTML-коду;

– форматування та колір тексту;
– наявність підозрілих посилань;
– результати перевірок SPF, DKIM та DMARC;
– інформація із зовнішніх чорних списків;
– статистичні характеристики повідомлення.

Правила можуть мати різний рівень складності. Найпростіші з них перевіряють наявність окремих слів або шаблонів тексту. Більш складні правила виконують мережеві перевірки, аналізують репутацію відправника або звертаються до зовнішніх баз даних відомих джерел спаму.

Важливою перевагою SpamAssassin є його висока гнучкість. Система підтримує багаторівневе налаштування параметрів як для окремих користувачів, так і для всієї організації. Адміністратор може змінювати порогове значення класифікації, налаштовувати ваги окремих правил, створювати власні критерії виявлення спаму та підключати додаткові модулі аналізу.

Завдяки можливості тонкого налаштування SpamAssassin може адаптуватися до специфіки поштового трафіку конкретної організації. Практика експлуатації показує, що підвищення точності фільтрації часто досягається шляхом аналізу помилково класифікованих повідомлень та коригування ваг тих правил, які спричинили неправильне рішення системи.

Попри високу ефективність, правила-орієнтований підхід має певні обмеження. Зокрема, підтримка великої кількості правил потребує постійного оновлення, а спамери можуть змінювати структуру повідомлень для обходу відомих критеріїв виявлення. Саме тому сучасні системи захисту електронної пошти дедалі частіше поєднують правила-орієнтовані методи з алгоритмами машинного навчання, серед яких особливе місце займають байєсовські класифікатори. Використання статистичних моделей дозволяє автоматично адаптувати систему до нових типів спаму та підвищувати точність класифікації повідомлень.

4 Висновки

У статті проведено комплексне теоретичне та аналітичне дослідження, присвячене взаємозв'язку між технологіями автоматичного моделювання природної мови та сучасними інфраструктурними й інтелектуальними методами захисту інформаційно-комунікаційних систем від небажаної кореспонденції (спаму).

За результатами проведеної роботи сформульовано такі основні висновки:

1. Ефективність нейромережевих архітектур у генерації тексту. Досліджено структуру та принципи функціонування рекурентних нейронних мереж (RNN). Обґрунтовано, що класичні RNN мають обмеження при роботі з довгими текстовими послідовностями через проблему затухання та вибуху градієнтів. Проаналізовано архітектуру моделей довгої короткочасної пам’яті (LSTM), яка за рахунок використання спеціальних комірок пам’яті та системи воріт (забування, вхідних і вихідних) успішно вирішує цю проблему. Це дозволяє моделювати складні довгострокові контекстні залежності й автоматично

генерувати високореалістичні тексти, що, з одного боку, розвиває технології NLP, а з іншого – створює нові виклики для систем інформаційної безпеки.

2. Еволюція загрозоутворюючих факторів (спаму). Систематизовано сучасні категорії спаму, серед яких особливу загрозу становлять фішинг, підробка адрес відправників (spoofing), масові розсилки комерційного характеру та повідомлення, що містять шкідливе програмне забезпечення. Констатовано, що автоматизація створення унікального спам-контенту за допомогою штучного інтелекту вимагає відходу від застарілих сигнатурних методів фільтрації.

3. Роль інфраструктурних методів автентифікації пошти. Розглянуто комплекс криптографічних та технічних засобів захисту на рівні поштових серверів: протоколи SPF, DKIM та DMARC. Встановлено, що їхнє спільне застосування є обов'язковим першим рубежем оборони, оскільки вони дозволяють жорстко верифікувати правомірність відправника та цілісність листа, ефективно блокуючи спроби імперсоналізації та підробки доменів. Також визначено специфіку допоміжних засобів, таких як «сірі списки» (greylisting), що базуються на поведінкових особливостях спам-ботів.

4. Переваги інтелектуального аналізу вмісту (контенту). Доведено, що для протидії спам-повідомленням, які пройшли інфраструктурні фільтри, найбільш ефективним є застосування наївного байєсівського класифікатора. Математичний апарат теорії ймовірностей дозволяє динамічно оцінювати частоту появи певних слів у легітимній та шкідливій кореспонденції, забезпечуючи високу точність адаптивної фільтрації тексту навіть в умовах постійної зміни тактик зловмисників.

Практична цінність та перспективи подальших досліджень. Матеріали дослідження показують, що максимальний рівень захисту сучасних інформаційних систем від автоматизованого спаму досягається лише за умови побудови ешелонованої системи безпеки. Вона має поєднувати інфраструктурну автентифікацію (SPF/DKIM/DMARC) та глибокий інтелектуальний аналіз вмісту повідомлень на основі статистичних методів і машинного навчання. Подальші дослідження доцільно спрямувати на розробку комбінованих алгоритмів, що інтегрують архітектури трансформерів (на кшталт BERT чи GPT) безпосередньо у контури систем фільтрації для виявлення контекстуально складного спаму в режимі реального часу.

Список використаних джерел

1. Understanding LSTM Networks [Електронний ресурс] – Режим доступу до ресурсу: http://colah.github.io/posts/2015-08-UnderstandingLSTMs/
2. Rosmunie Y. B. Should we be concerned with spam emails? A look at its impacts and implications [Електронний ресурс] / Y. B. Rosmunie, H. Husnayati – Режим доступу до ресурсу: https://ieeexplore.ieee.org/document/6518877.
3. Roney A. Cyber crimes: Threats and protection [Електронний ресурс] / A. Roney, A. Al Hajj, K. Al Ruqeishi – Режим доступу до ресурсу: https://ieeexplore.ieee.org/document/5508568.

4. Dhinakaran C. "Reminder: please update your details": Phishing Trends [Електронний ресурс] / C. Dhinakaran, J. Kwang Lee. – 2019. – Режим доступу до ресурсу: https://ieeexplore.ieee.org/document/5383991.
5. Postel J. B. Simple mail transfer protocol [Електронний ресурс] / Jonathan B. Postel. – 1982. – Режим доступу до ресурсу: https://datatracker.ietf.org/doc/html/rfc821.
6. Kitterman S. Sender Policy Framework (SPF) [Електронний ресурс] / S. Kitterman. – 2014. – Режим доступу до ресурсу: https://datatracker.ietf.org/doc/html/rfc720.
8. Crocker D. DomainKeys Identified Mail (DKIM) Signatures [Електронний ресурс] / D. Crocker, T. Hansen, M. Kucherawy. – 2011. – Режим доступу до ресурсу: https://datatracker.ietf.org/doc/html/rfc6376.
9. Kucherawy M. Domain-based Message Authentication, Reporting, and Conformance (DMARC) [Електронний ресурс] / M. Kucherawy, E. Zwicky. – 2015. – Режим доступу до ресурсу: https://datatracker.ietf.org/doc/html/rfc7489.
10. Artificial hygiene: a critical step towards safety from email viruses [Електронний ресурс] / A.Talukder, V. Rao, V. Kapoor, D. Sharma. – 2005. – Режим доступу до ресурсу: https://ieeexplore.ieee.org/document/1497801
11. imensions of Cyber Attack [Електронний ресурс] / R.A Gandhi, A. Sharma, W. Mahoney, W. Sousan. – 2011. – Режим доступу до ресурсу: https://www.researchgate.net/publication/224223630_Dimensions_of_Cyber-Attacks_Cultural_Social_Economic_and_Political. 11. Prachi G. J. A Survey on Email Spam Types and Spam Filtering Techniques / G. J. Prachi, K. P. R. // International Journal of Engineering Research & Technology. – 2014. – №2278. – С. 2. – Режим доступу до ресурсу: https://www.ijert.org/research/a-survey-on-email-spam-types-and-spam-filtering-techniques-IJERTV3IS031626.pdf
12. Kucherawy M. mail Authentication Status Codes [Електронний ресурс] / M. Kucherawy. – 2014. – Режим доступу до ресурсу: https://datatracker.ietf.org/doc/html/rfc7372.
13. Fagerland V. Automatic Analysis of Scam Emails [Електронний ресурс] / Vegard Fagerland. – 2017. – Режим доступу до ресурсу: https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2461337/17721_FULLTEXT.pdf
14. Email Authentication with Sender ID [Електронний ресурс] – Режим доступу до ресурсу: https://adventuresinsecurity.com/blog/?p=78
15. Terry Zink. Where email authentication falls flat at stopping phishing – impersonation attacks using display tricks, 2016. URL https://blogs.msdn.microsoft.com/tzink/2016/12/06/ where-email-authentication-falls-flat-at-stopping-phishing-impersonation-attacks-using-display-tricks/.
16. Zink T. Where email authentication falls flat at stopping phishing - impersonation attacks using display tricks [Електронний ресурс] / Terry Zink. – 2016. – Режим доступу до ресурсу: https://docs.microsoft.com/en-us/archive/blogs/tzink/where-

email-authentication-falls-flat-at-stopping-phishing-impersonation-attacks-using-display-tricks.
17. What is DMARC? [Електронний ресурс]. – 2017. – Режим доступу до ресурсу: https://dmarc.org/.
18. Research of a Spam Filtering Algorithm Based on Naive Bayes and AIS [Електронний ресурс] / Q.Luo, B. Liu, J. Yan, Z. He. – 2010. – Режим доступу до ресурсу: https://ieeexplore.ieee.org/document/5709036.
19. SpamAssassin [Електронний ресурс] – Режим доступу до ресурсу: https://spamassassin.apache.org.

DOI 10.70286/EOSS-20.07.2026.007.128-136

ENHANCING SQUEEZENET-BASED METAL DEFECT CLASSIFICATION UNDER DATA SCARCITY VIA DCGAN-DRIVEN SYNTHESIS

Vasko Artem
postgraduate student
Computer Science Department
Olesya Honchara Dnipro National University, Ukraine

Abstract. Automated visual inspection of metal surfaces relies on convolutional neural networks that require large labelled datasets, yet in real manufacturing lines defect images are scarce and unevenly distributed across classes. This study investigates whether synthetic samples produced by a Deep Convolutional Generative Adversarial Network (DCGAN) can compensate for data scarcity when training a lightweight SqueezeNet classifier. Class-specific DCGANs were trained on limited real images from the NEU surface-defect benchmark and used to expand the training pool before SqueezeNet fine-tuning. Across four scarcity levels the DCGAN-augmented model consistently outperformed both the real-only baseline and classic geometric augmentation, raising test accuracy from 71.3% to 84.6% in the most severe scarcity setting (30 images per class) while keeping the model under 1.25 million parameters.

Keywords: metal surface defect, SqueezeNet, DCGAN, generative adversarial network, data augmentation, data scarcity, convolutional neural network.

Introduction. Surface-defect inspection is a decisive quality-control stage in metallurgy, automotive and machine-building industries, where undetected cracks, scratches or inclusions propagate into costly downstream failures. Over the past decade manual inspection has been progressively replaced by deep convolutional neural networks, which achieve human-level recognition of steel-strip defects when trained on sufficiently large datasets (Krizhevsky et al., 2012; He et al., 2016). In practice,

Proceedings of the 5th International Scientific
and Practical Conference
"Scientific Research: Modern Innovations and Future Perspectives"
July 20-22, 2026
Montreal, Canada

Contact details of the organizing committee:
European Open Science Space
E-mail: info@eoss-conf.com
URL: https://www.eoss-conf.com/

EUROPEAN OPEN
SCIENCE SPACE