КОМП'ЮТЕРНІ НАУКИ

УДК 004.8:004.65:004.728.8:004.056
DOI: https://doi.org/10.17721/ISTS.2025.9.74-80

Віктор МОРОЗОВ, канд. техн. наук, проф.
ORCID ID: 0000-0001-7946-0832
e-mail: viktor.morozov@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

Владислав ДЕЙНЕГА, асп.
ORCID ID: 0009-0008-3123-2302
e-mail: thedynkan@knu.ua
Київський національний університет імені Тараса Шевченка, Київ, Україна

ВИЯВЛЕННЯ СПАМУ В ТЕКСТОВИХ ПОВІДОМЛЕННЯХ ІЗ ВИКОРИСТАННЯМ ЛОГІСТИЧНОЇ РЕГРЕСІЇ НА БАЗІ ГРАДІЄНТНОГО СПУСКУ

Вступ. *Зі зростанням обсягів електронного листування проблема фільтрації спаму набуває все більшої актуальності. За даними статистичних досліджень, спам становить значну частку глобального поштового трафіка, що створює ризики як для безпеки, так і для ефективності електронної комунікації. У цьому контексті особливого значення набувають методи оброблення природної мови (NLP) та машинного навчання. Метою цієї роботи є побудова моделі для класифікації електронних повідомлень на спам і не спам, із використанням логістичної регресії, реалізованої через градієнтний спуск, у поєднанні з методами оброблення текстових даних.*

Методи. *Для навчання моделі використано датасет, що містить понад 5000 електронних повідомлень, мічених як спам або не спам. Дані було попередньо очищено з видаленням шумових компонентів: пунктуації, цифр, стоп-слів, коротких слів, а також застосовано лематизацію. Тексти перетворено на числову форму за допомогою TF-IDF векторизації із L2-нормалізацією. Для боротьби з дисбалансом між класами застосовано метод SMOTE. Навчання моделі здійснювалось за класичною схемою градієнтного спуску з використанням сигмоїдної функції активації та логарифмічної функції втрат.*

Результати. *Побудована модель досягла високих результатів на тестовій вибірці: загальна точність становила 98 %, f1-score для класу спам – 0.92, а для не спаму – 0.99. Значення recall для спаму дорівнювало 0.90, що свідчить про здатність моделі виявляти більшість небажаних повідомлень без надмірних помилкових спрацьовувань. Баланс precision і recall також підтверджується макросереднім і зваженим середнім f1-показником понад 0.96.*

Висновки. *Результати дослідження засвідчили ефективність поєднання логістичної регресії, градієнтного спуску та текстового препроцесингу для задачі класифікації спаму навіть за умов дисбалансованих даних. Запропонований підхід є ефективним й інтерпретованим, що робить його придатним для практичного застосування в системах фільтрації електронної пошти.*

Ключові слова: *машинне навчання, оброблення природної мови, градієнтна оптимізація, фільтрація електронної пошти, текстовий препроцесинг, класифікація.*

Вступ

В сучасному цифровому просторі електронна пошта залишається одним із ключових засобів комунікації, що використовується як приватними користувачами, так і комерційними структурами. Водночас цей канал щодня генерує колосальні обсяги повідомлень – за оцінками, понад 330 млрд електронних листів щодня надсилається по всьому світу. Серед них, за статистикою, спам-повідомлення становлять приблизно 46 % загального трафіка, що свідчить про стабільно високу частку небажаних листів у комунікаційній інфраструктурі. Отже, щодня у світі поширюється понад 150 млрд спам-листів. Такі повідомлення можуть містити рекламний або шахрайський контент, шкідливі вкладення або фішингові посилання, що створює не лише інформаційне перевантаження для користувача, а й потенційні ризики для безпеки даних. Ці показники підтверджують і масштабність, і сталість проблеми, що актуалізує потребу в розробленні ефективних та адаптивних систем автоматичної фільтрації небажаної електронної кореспонденції.

Проблема автоматичного виявлення спаму полягає в необхідності точно розпізнавати небажані повідомлення на основі їхнього текстового вмісту. Прості методи фільтрації вже недостатньо ефективні, тому зростає потреба в алгоритмах, здатних аналізувати зміст повідомлень і виявляти приховані шаблони. У зв'язку із цим зростає актуальність застосування методів оброблення природної мови (NLP) та машинного навчання, які дозволяють автоматично виявляти закономірності у текстах на основі попередньо навчених моделей.

Серед найбільш інтерпретованих і водночас ефективних моделей для задач бінарної класифікації – логістична регресія. Її особливість полягає в тому, що вона дає змогу отримувати ймовірнісні передбачення та не потребує надмірної кількості ресурсів. У цьому дослідженні логістична регресія реалізується через градієнтний спуск – класичний метод оптимізації, що дозволяє поступово оновлювати ваги моделі з урахуванням похибки.

Важливою передумовою якісного навчання є глибокий текстовий препроцесинг, який охоплює очищення повідомлень від пунктуації, чисел, надлишкових пробілів, стоп-слів і коротких слів, а також лематизацію. Саме цей етап дозволяє зменшити шум у даних і виділити лише ті слова, що несуть змістове навантаження. Далі тексти перетворено на числову форму за допомогою TF-IDF векторизації, що дозволяє враховувати не тільки частоту терміна, а і його інформативність у межах усього корпусу. У векторизації застосовують L2-нормалізацію, яка робить ознаки порівнюваними за масштабом і покращує роботу оптимізаційного алгоритму.

Окремим викликом у роботі з реальними даними є дисбаланс між класами, коли повідомлення певного типу (зазвичай – не спам) значно переважають. У таких

ISSN 2707-1758

умовах важливо застосовувати методи балансування, які дають змогу зробити навчання моделі чутливішим до менш представленого класу. У нашому дослідженні для цього використано підхід синтетичного збільшення вибірки (SMOTE – Synthetic Minority Over-sampling Technique), який дозволяє зберегти обсяг даних класу більшості, водночас підвищуючи представленість меншості.

Метою дослідження є розроблення моделі для класифікації електронних повідомлень на спам і не спам на основі логістичної регресії, оптимізованої методом градієнтного спуску, із залученням методів оброблення текстових даних.

Огляд літератури. Дослідження "Spam Statistics 2025: New Data on Junk Email, AI Scams & Phishing" (Spam Statistics, 2025) надає актуальні статистичні дані щодо спаму, фішингу та пов'язаних із ними загроз у сфері електронної пошти. Згідно з отриманими результатами, у 2023 р. щодня надсилалося близько 160 млрд спам-листів, що становило приблизно 46 % загального обсягу електронної пошти. Це свідчить про значну частку небажаних повідомлень у глобальному електронному трафіку. Опитування також показало, що 96,8 % користувачів отримували спам-повідомлення в тій чи іншій формі. Найпоширенішими темами спаму були: призи та розіграші (36,7 %), пропозиції роботи (36,3 %) та банківські повідомлення (34,6 %). Ці дані підкреслюють, що спамери часто використовують теми, які можуть привернути увагу користувачів і спонукати їх до взаємодії з повідомленням. Крім того, дослідження виявило, що 68,8 % респондентів повідомили про негативний вплив спаму та фішингових повідомлень на їхній психічний стан, що свідчить про психологічні наслідки таких загроз. Також зазначено, що фінансові установи є найчастішими цілями фішингових атак, отримуючи 27,7 % таких повідомлень. Це дослідження надає цінну інформацію для розуміння масштабів та еволюції спаму, а також підкреслює необхідність розроблення ефективних методів виявлення та фільтрації небажаних повідомлень.

Публікація автора Jyothiikaa Moorthy (2025) підтверджує ключові висновки попереднього дослідження щодо масштабності проблеми спаму, зафіксувавши його частку на рівні 45,6 % у 2023 р. з подальшим зростанням до 46,8 % у грудні 2024 р. Також проінформовано, що майже 96 % фішингових атак здійснюють за допомогою електронної пошти. Окрім кількісних даних, звіт також доповнює попереднє джерело деталізацією тематики спаму: найпоширенішими є маркетингові повідомлення (36 %), контент для дорослих (31,7 %) і повідомлення, пов'язані з фінансовими питаннями (26,5 %). Така типізація розширює уявлення про природу спаму й підкреслює потребу в адаптивних моделях виявлення залежно від його змісту.

У статті (Khanday et al., 2021) розглянуто використання логістичної регресії для класифікації електронних листів на спам і не спам. Автори підкреслюють універсальність логістичної регресії як одного з найкращих методів для класифікації електронних листів. Вказана стаття може бути корисною для дослідження, оскільки вона демонструє практичне застосування логістичної регресії у задачах класифікації спаму.

У роботі (Zhang et al., 2019) логістичну регресію застосовують до задачі класифікації у фінансовому секторі за умов значного дисбалансу класів. Автори демонструють, що навіть у таких складних умовах, із правильною адаптацією, цей метод залишається ефективним. Цей підхід є релевантним до нашого дослідження, де модель логістичної регресії також використовується в задачі з дисбалансованими даними, і доводить, що у разі належної структури оброблення даних модель здатна зберігати високу чутливість до меншості класу.

У статті (Khlevna, & Deineha, 2023) також розглянуто використання логістичної регресії для класифікації шахрайства, де так само можна бачити важливість попереднього оброблення даних, яке є одним із важливих елементів роботи.

У дослідженні (Rakhmanov, 2020) на підставі аналізу тональності в освітньому середовищі проведено порівняння методів векторизації та класифікації на основі понад 52 000 текстових коментарів. Автори показали, що Tf-IDF – Term Frequency–Inverse Document Frequency – перевершує CountVectorizer за точністю класифікації, що підтверджує доцільність використання саме цього підходу у задачах текстової класифікації. Отримані результати узгоджуються з підходом, застосованим у нашій роботі, де TF-IDF також використовується як основна техніка векторного подання тексту.

Зауважимо, що в дослідженні (Papageorgiou, Economou, & Bersimis, 2024) з класифікації бізнес-електронної пошти розглянуто ті самі ключові етапи, що і в нашій роботі: препроцесинг, векторизація тексту та класифікація email-повідомлень, що підтверджує актуальність обраного підходу.

Mohammed, N., Mouhajir, M., & Yassine S. (2023) описали логістичну регресію з градієнтним спуском, що прямо відповідає підходу, застосованому і в нашому дослідженні. Обраний метод навчання є релевантним і демонструє практичну цінність градієнтного спуску для задач класифікації.

Методи

Попереднє оброблення текстових даних є критично важливою складовою будь-якого завдання машинного навчання, пов'язаного з обробленням природної мови, оскільки саме на цьому етапі формується чисте, лаконічне й інформативне представлення вхідного тексту, яке згодом буде перетворене у вектори для навчання моделі. У межах цього дослідження реалізовано поетапний підхід до оброблення текстів, який охоплює лінгвістичне очищення та структуроване скорочення шуму в даних.

На першому етапі всі електронні повідомлення було перетворено у нижній регістр для уникнення зайвої варіативності слів, що мають однакову семантику, але різне написання. Це дозволяє уникнути дублювання слів, які відрізняються лише регістром, наприклад "Free" та "free". Далі застосовано регулярні вирази для видалення чисел, знаків пунктуації, символів нового рядка та надлишкових пробілів, що часто не несуть смислового навантаження у задачах класифікації спаму.

Після цього текст розбивався на токени – окремі слова – за допомогою стандартного токенізатора бібліотеки nltk. Кожен токен перевірявся на наявність у стандартному списку англійських стоп-слів, і непотрібні функціональні слова (як-от "the", "is", "and") були виключені. Це дозволяє моделі зосередитись на більш значущих елементах вхідного тексту.

Наступним етапом була лематизація, тобто приведення кожного токена до його початкової граматичної форми. Для цього використовували лематизацію з урахуванням частини мови (POS-теги),

ISSN 2707-1758

що дозволяє досягти кращої точності порівняно з простим зведенням до базової форми (рис. 1).

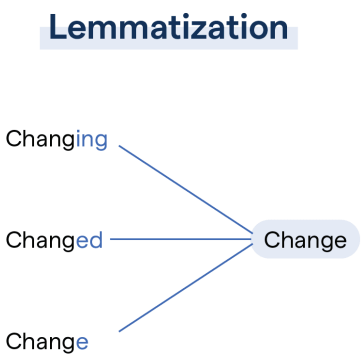

Рис. 1. Ілюстрація процесу лематизації

Додатково запроваджено фільтрацію коротких слів, зокрема тих, що мають довжину менше трьох символів. Це дозволяє зменшити кількість шумових токенів, які часто мають низьку інформативність.

В результаті оброблення кожне повідомлення перетворено в очищений і лематизований текст, з якого видалено шумові та малозначущі елементи. Такий підхід дозволяє зберегти змістовну суть повідомлення та забезпечує узгоджене представлення всіх текстів у корпусі.

Цей етап препроцесингу є критично важливим для коректної роботи моделі логістичної регресії, яка не має внутрішнього механізму розуміння тексту, і працює винятково із числовим представленням вхідних даних. У нашій роботі завдяки структурованому препроцесингу вдалося підготувати тексти до векторизації через TF-IDF, забезпечивши як однорідність корпусу, так і збереження релевантної лексичної інформації, необхідної для успішного навчання моделі.

Після завершення етапу текстового препроцесингу всі вхідні повідомлення мають вигляд очищених рядків, що складаються з лематизованих слів без пунктуації, чисел, стоп-слів і зайвих символів. У такому вигляді текст все ще залишається неструктурованим і непридатним до оброблення алгоритмом машинного навчання, оскільки модель працює виключно із числовими значеннями. Тому наступним важливим етапом є перетворення текстових документів у векторну форму – процес, відомий як векторизація.

Після завершення препроцесингу всі очищені повідомлення було розділено на навчальну та тестову вибірки у пропорції 70:30. Такий розподіл дозволяє забезпечити об'єктивне оцінювання моделі на нових, раніше не бачених даних, і запобігає переоцінюванню її точності.

У цьому дослідженні для представлення текстів у вигляді числових векторів використано метод TF-IDF, який враховує не лише наявність слова у повідомленні, а і його важливість у межах усього корпусу. Цей підхід дозволяє приділяти більше уваги тим словам, які є характерними для окремих повідомлень, і приглушувати вагу слів, що часто трапляються повсюдно. Наприклад, якщо слово часто зустрічається в усіх повідомленнях, то його інформативність вважається нижчою.

Метод TF-IDF поєднує два складники. Перший – TF (term frequency) – відображає частоту появи терміна у конкретному документі. Другий – IDF (inverse document frequency) – зменшує вагу слів, що зустрічаються у багатьох документах, і підвищує значущість тих, які є рідкісними, але характерними. Добуток цих двох значень формує остаточну вагу кожного слова в повідомленні. У результаті векторизації кожен документ отримує числове представлення, яке відображає важливість використаних у ньому термінів. Формула для обчислення TF-IDF:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right), \quad (1)$$

де $w_{i,j}$ – вага (важливість) терміна i в документі j, $tf_{i,j}$ – частота терміна i в документі j (Term Frequency), N – загальна кількість документів, df_i – кількість документів, у яких зустрічається термін i (Document Frequency).

Для реалізації TF-IDF векторизації використано TfidfVectorizer із бібліотеки scikit-learn. Попередньо очищені повідомлення спочатку були подані до методу fit_transform() для навчального набору, що дозволило побудувати словник усіх термінів та обчислити відповідні ваги. Тестова вибірка векторизується за допомогою методу transform() – виключно на основі словника, отриманого з навчальних даних. Такий підхід запобігає витіканню інформації з тестової вибірки до моделі та забезпечує коректність оцінювання якості класифікації.

У процесі векторизації була застосована L2-нормалізація, яка перетворює кожен текстовий вектор так, щоб його довжина у просторі ознак дорівнювала одиниці. Це сприяє уніфікації масштабів ознак і забезпечує коректну роботу оптимізатора. Нормалізація є важливою складовою підготовки даних для ефективного застосування градієнтного спуску.

У результаті етапу векторизації весь текстовий корпус було перетворено у числову форму, придатну для навчання моделі логістичної регресії. Отримані вектори містять зважену інформацію про частотність і значущість слів у контексті задачі, що забезпечує якісну основу для подальшої класифікації повідомлень як спам або не спам. Цей підхід поєднує простоту реалізації та високу інформативність, що робить TF-IDF ефективним інструментом у задачах оброблення природної мови.

Однією з характерних особливостей задач класифікації у сфері електронної пошти є суттєвий дисбаланс класів. У типових наборах даних спостерігається значна перевага одного класу – наприклад, легітимних повідомлень – над іншим, яким часто є спам. У нашому дослідженні такий дисбаланс початково становив близько 13 % і до 87 %, а це створює загрозу, що модель навчиться надавати перевагу домінуючому класу, ігноруючи менш представлений.

Щоб компенсувати цей ефект, застосовано метод штучного збільшення вибірки меншості – SMOTE. Вказаний метод SMOTE є одним із найпопулярніших підходів до oversampling і полягає у синтетичному створенні нових прикладів меншості на основі наявних. На відміну від простого дублювання, SMOTE генерує нові точки інтерполяцією між наявними прикладами того самого класу. Це дозволяє моделі бачити більшу варіативність прикладів меншості і краще навчатися на її представленнях (рис. 2).

SMOTE застосовано лише до навчальної вибірки (X_train і y_train). Тестова вибірка залишалася незмінною для забезпечення об'єктивності та репрезентативності оцінки продуктивності моделі. Такий підхід узгоджується із загальноприйнятими практиками машинного навчання і запобігає "витіканню інформації" з тестових даних до процесу навчання.

ISSN 2707-1758

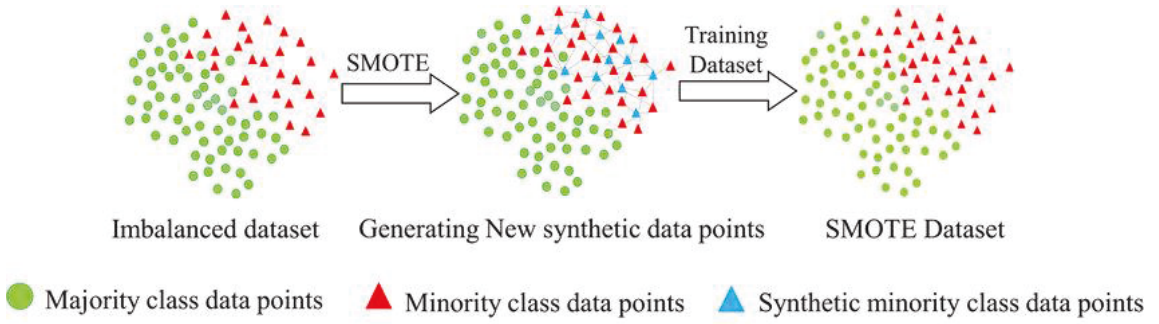

Рис. 2. Візуалізація процесу застосування методу SMOTE

Перед застосуванням SMOTE усі вхідні тексти вже були подані у векторизованому TF-IDF вигляді, а отже, представлені у вигляді числових ознак, що є необхідною умовою для його коректної роботи. SMOTE не працює безпосередньо з текстами, а лише із числовим простором ознак – тому його місце у конвеєрі оброблення даних розташовано після векторизації, але до навчання моделі.

Для реалізації методу SMOTE використовувався модуль SMOTE з бібліотеки imbalanced-learn. Алгоритм базується на підході, який використовує *k*-ближчих сусідів для генерації нових синтетичних зразків класу меншості, зберігаючи його основну структуру у векторному просторі.

Рішення обмежитися лише SMOTE без використання undersampling пояснюється прагненням зберегти всі наявні приклади класу більшості, оскільки видалення частини даних могло б призвести до втрати важливої інформації та зниження загальної стійкості моделі.

Після балансування класів за допомогою SMOTE навчальна вибірка стала симетричною, що дозволяє алгоритму логістичної регресії навчатись без упередженості до одного з класів. Отже, застосування SMOTE стало ключовим етапом у підготовці навчальних даних до моделювання і дозволило уникнути типових помилок, пов'язаних з ігноруванням меншості класу, а також забезпечити збалансовану основу для побудови ефективної класифікаційної моделі.

До векторного представлення повідомлень було додано допоміжну колонку зі значенням 1 для кожного прикладу. Це стандартний технічний прийом, який дозволяє моделі враховувати зміщення під час навчання.

Для розв'язання задачі бінарної класифікації в цьому дослідженні застосовується логістична регресія, яка дозволяє прогнозувати ймовірність належності об'єкта до одного з двох класів. На відміну від лінійної регресії, логістична модель обмежує вихід значенням у межах інтервалу [0;1], що дозволяє трактувати результат як імовірність позитивного класу – у нашому випадку, повідомлення, класифікованого як "спам".

Ключовим елементом логістичної регресії є сигмоїдна функція активації, яка перетворює лінійну комбінацію вхідних ознак у значення від 0 до 1:

$$\sigma(x) = \frac{1}{1+e^{-x}}, \quad (2)$$

де *x* – вхідне значення (лінійна комбінація вхідних ознак та їхніх ваг), *e* – число Ейлера.

Ваги визначають у процесі навчання моделі на навчальному наборі даних. Таким способом, кожному прикладу присвоюється ймовірність належності до позитивного класу. Остаточне рішення приймається шляхом порівняння з порогом (зазвичай, 0,5): якщо значення сигмоїди більше порога – прогнозується клас 1, в іншому разі – клас 0.

Навчання моделі здійснюється за допомогою градієнтного спуску – одного з найпоширеніших чисельних методів оптимізації. Ідея цього методу полягає у поступовому оновленні параметрів моделі в напрямку, який найшвидше зменшує значення функції втрат. У логістичній регресії такою функцією виступає логарифмічна функція втрат (Log*Loss*):

$$\mathrm{Log}Loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \log(p_i) + (1-y_i)\cdot\log(1-p_i)), \quad (3)$$

де *N* – кількість спостережень, y_i – фактичний бінарний результат (0 або 1) для *i*-го спостереження, p_i – прогнозована ймовірність того, що *i*-те спостереження належить до класу 1.

Оновлення ваг відбувається за формулою

$$\theta_{new} = \theta_{old} - \alpha \cdot \nabla J(\theta), \quad (4)$$

де θ – параметри моделі, α – швидкість навчання, а $\nabla J(\theta)$ – градієнт функції втрат $J(\theta)$.

Отже, під час кожної ітерації модель оновлює параметри у напрямку зменшення похибки. Процес продовжується доти, поки або зміни у функції втрат не стануть незначними, або не буде досягнуто заданої кількості ітерацій.

Важливим аспектом є початкова ініціалізація ваг, яка може здійснюватися нульовими або малими випадковими значеннями. Також критичним є вибір параметра α, оскільки надто велике значення може спричинити нестабільність, а занадто мале – уповільнити збіжність.

У підсумку градієнтний спуск дозволяє поступово налаштувати модель логістичної регресії для точного прогнозування класу спам / не спам на основі векторизованих текстових даних. Завдяки аналітичній простоті й математичній інтерпретованості цей підхід є придатним як для практичного використання, так і для глибокого контролю над навчанням моделі.

Оцінювання ефективності класифікаційної моделі є ключовим етапом, що дозволяє визначити її здатність узагальнювати закономірності та правильно розпізнавати об'єкти обох класів. У задачі виявлення спаму, де класи істотно нерівномірно представлені, особливо важливо враховувати не лише загальну точність, але й інші якісні характеристики моделі.

Основним інструментом оцінювання є матриця невідповідностей (confusion matrix), яка фіксує кількість правильно та неправильно класифікованих повідомлень кожного класу. У нашому випадку клас "спам" позначено як позитивний (positive class).

ISSN 2707-1758

Отже маємо:

- True Positives (TP) – спам-повідомлення, правильно класифіковані як спам;
- False Negatives (FN) – спам, помилково класифікований як не спам;
- False Positives (FP) – не спам, класифікований як спам;
- True Negatives (TN) – не спам, правильно класифікований як не спам.

На основі цієї матриці обчислюють основні метрики:

Accuracy – загальна частка правильних передбачень:

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP}. \quad (5)$$

Precision (точність) для класу "спам" показує, яку частку з усіх передбачених як спам повідомлень справді становить спам:

$$Precision = \frac{TP}{TP+FP}. \quad (6)$$

Recall (повнота) для класу "спам" визначає, яку частку зі справжніх спамів модель змогла правильно розпізнати:

$$Recall = \frac{TP}{TP+FN}. \quad (7)$$

F1-score є гармонічним середнім між точністю та повнотою:

$$F1 = 2\frac{Precision*Recall}{Precision+Recall}. \quad (8)$$

У підсумку багатовимірна система оцінювання дозволяє всебічно проаналізувати роботу класифікатора. Метрики precision, recall та f1-score дають можливість окремо оцінити здатність моделі виявляти спам та уникати хибних спрацьовувань для не спаму. Це має вирішальне значення для реальних сценаріїв автоматичного фільтрування електронної пошти, де помилки можуть призводити як до втрати важливих повідомлень, так і до пропуску небажаного контенту.

Результати

Для проведення експериментального дослідження використано датасет The Enron Email Dataset, доступний на платформі Kaggle (2025). Цей набір містить 5 157 електронних повідомлень, кожне з яких має текстову частину та мітку, що вказує на його належність до одного з двох класів: "ham" (звичайне повідомлення) або "spam" (небажане повідомлення).

Дані представлені у вигляді двох колонок:

- Message – текст електронного листа (англійською мовою).
- Category – мітка класу (ham або spam), що використовується як цільова змінна для класифікації.

Розподіл класів є нерівномірним:

- 87 % повідомлень належать до класу ham (не спам),
- 13 % – до класу spam.

Такий дисбаланс типово спостерігається в реальних електронних скриньках і створює додаткові складнощі для навчання моделей, оскільки алгоритми машинного навчання мають схильність до домінуючого класу. Саме тому під час оброблення даних було використано методи, спрямовані на компенсацію цього дисбалансу, зокрема і SMOTE.

Самі повідомлення у датасеті мають довжину від одного до кількох речень і містять як звичайну, так і рекламну лексику. Спам-повідомлення, зазвичай, включають фрази на кшталт "win", "free entry", "urgent", які характерні для фішингових або рекламних листів. Натомість повідомлення класу "ham" охоплюють щоденне листування, жарти, короткі особисті повідомлення тощо.

Повідомлення не мають структурованих полів, таких як тема чи адресат, і розглядаються виключно як неструктурований текст, що робить задачу лінгвістичною. Це дозволяє сфокусуватись на аналізі змісту повідомлень і якісному препроцесингу, без залежності від метаданих.

Отже, вибраний набір даних є репрезентативним прикладом задачі фільтрації електронної пошти на рівні змісту повідомлень і дозволяє моделювати класифікаційні алгоритми з орієнтацією на реальні сценарії застосування.

Розроблена модель логістичної регресії, навчена за допомогою градієнтного спуску, продемонструвала високі показники якості при класифікації електронних повідомлень на дві категорії: "спам" і "не спам" (рис. 3). Для оцінювання її ефективності використовували тестову вибірку, яка була попередньо відокремлена від навчальної на етапі препроцесингу. Таким способом, було забезпечено об'єктивність вимірювання результатів і виключено можливість витікання інформації з навчального етапу.

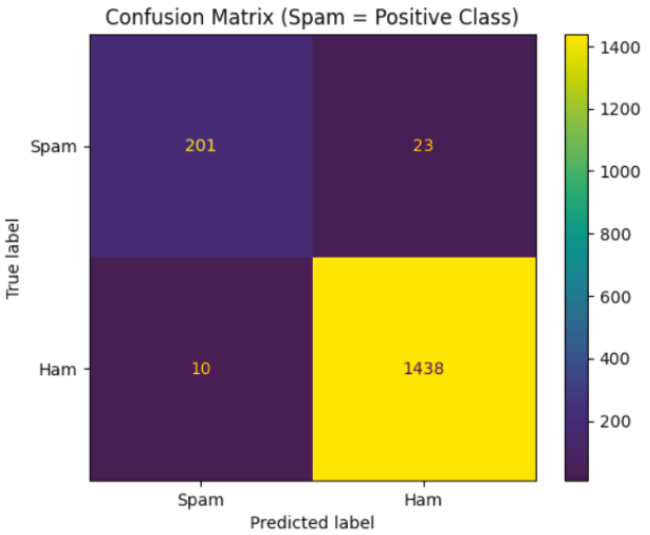

Рис. 3. Матриця невідповідностей для тестової вибірки (Spam – позитивний клас)

ISSN 2707-1758

Результати класифікації представлено у вигляді матриці невідповідностей (confusion matrix), яка відображає кількість правильно та неправильно класифікованих повідомлень обох класів (рис. 4):

- 201 спам-повідомлення були правильно класифіковані як спам (True Positives),
- 23 спам-листи були помилково розпізнані як не спам (False Negatives),
- 10 звичайних повідомлень були помилково віднесені до спаму (False Positives),
- 1438 не спамів були коректно класифіковані як не спам (True Negatives).

Загальна точність класифікації (accuracy) становила 98 %, що свідчить про те, що абсолютна більшість передбачень моделі були правильними. Проте під час оцінювання якості класифікації в умовах початкового дисбалансу класів важливо не обмежуватися лише загальною точністю. Саме тому проаналізовано глибші метрики: precision, recall і f1-score для кожного з класів окремо.

	precision	recall	f1-score	support
Ham	0.98	0.99	0.99	1448
Spam	0.95	0.90	0.92	224
accuracy			0.98	1672
macro avg	0.97	0.95	0.96	1672
weighted avg	0.98	0.98	0.98	1672

Рис. 4. Метрики оцінювання моделі логістичної регресії для тестової вибірки

Для класу "Ham", який є чисельнішим у датасеті, модель досягла таких результатів:

- Precision – 0.98: лише 2 % повідомлень, класифікованих як не спам, насправді були спамом.
- Recall – 0.99: з усіх реальних повідомлень, які не є спамом, 99 % були правильно виявлені.
- F1-score – 0.99: гармонійний баланс між точністю і повнотою.

Для класу "Spam", який є менш представленим:

- Precision – 0.95: з усіх передбачень класу спам, 95 % виявилися правильними.
- Recall – 0.90: модель виявила 90 % усіх реальних спамів у тестовій вибірці.
- F1-score – 0.92: значення, що підтверджує добре поєднання точності та виявлення спаму.

Macro average F1-score становить 0.96, що демонструє збалансовану продуктивність моделі для обох класів незалежно від їхньої частки в наборі даних. Weighted average F1-score дорівнює 0.98, що також підтверджує надійність моделі у практичному застосуванні.

Найбільший виклик традиційно полягає у виявленні меншості класу – спаму. Незважаючи на те, що Recall для нього дещо нижчий (0.90), модель продемонструвала здатність правильно класифікувати більшість спам-повідомлень і водночас не плутати їх із легітимними листами, що важливо для користувацького досвіду.

Отже, модель логістичної регресії на основі градієнтного спуску, у поєднанні з ретельним препроцесингом і методами балансування вибірки, продемонструвала високу ефективність у класифікації текстових електронних повідомлень, зокрема й у виявленні спаму в умовах обмежених і незбалансованих даних.

Дискусія і висновки

Результати дослідження підтверджують, що логістична регресія, реалізована з використанням градієнтного спуску, є ефективним і водночас інтерпретованим інструментом для класифікації текстових повідомлень на спам і не спам. Досягнута точність класифікації на рівні 98 % свідчить про високу здатність моделі до узагальнення навіть за умов обмеженого обсягу навчальних даних і наявного дисбалансу між класами.

Одним із ключових чинників, що вплинув на якість моделі, є етап підготовки текстових даних. Ретельне виконання препроцесингу – включаючи видалення пунктуації та чисел, фільтрацію стоп-слів, лематизацію та відсікання коротких токенів – дозволило перетворити необроблений текст на структуроване представлення з мінімумом шуму. Це створило необхідні передумови для ефективної векторизації та подальшого навчання.

Застосування TF-IDF векторизації забезпечило зважене врахування і частоти термінів, і їхньої релевантності в межах корпусу. Такий підхід дозволяє зосередитись на тих словах, що мають інформативну цінність, та зменшити вплив часто вживаних, але малозначущих слів. Важливою деталлю є використання L2-нормалізації, що вирівнює довжини векторів і сприяє стабільності та коректності сходження градієнтного алгоритму.

Оскільки початкові дані мали суттєвий дисбаланс між класами (87 % не спаму та 13 % спаму), особливу роль відіграло застосування методу SMOTE, що дозволив штучно збалансувати вибірку генерацією синтетичних прикладів меншості. Такий підхід дав змогу моделі побачити більше варіацій спам-повідомлень і підвищити recall без втрати точності.

Модель продемонструвала сильні показники як для переважаючого класу ("не спам"), так і для менш представленого класу ("спам"). Для класу спаму значення precision склало 0.95, recall – 0.90, а f1-score – 0.92, що свідчить про ефективність виявлення небажаних повідомлень при мінімальній кількості помилок. Для класу не спаму ці показники ще вищі: precision – 0.98, recall – 0.99, f1-score – 0.99. Макросереднє значення F1-метрики (0.96) і зважене середнє (0.98) підтверджують збалансованість моделі за всіма метриками, що важливо в умовах початкового перекосу у вибірці. Це означає, що модель не лише правильно розпізнає більшість повідомлень, але й зберігає стабільну якість на обох класах.

Загалом, результати підтверджують практичну придатність обраного підходу до автоматичної класифікації електронних листів. Поєднання логістичної регресії, градієнтного спуску, грамотно

ISSN 2707-1758

організованого препроцесингу й балансування вибірки дало змогу досягти високої точності при збереженні прозорості та керованості всієї моделі.

Внесок авторів: Віктор Морозов – концептуалізація; методологія; аналіз джерел, підготування огляду літератури або теоретичних засад дослідження; Владислав Дейнега – збір емпіричних даних та їхня валідація; емпіричне дослідження.

Джерела фінансування. Це дослідження не отримало жодного гранту від фінансової установи в державному, комерційному або некомерційному секторах.

Список використаних джерел

Jyothiikaa Moorthy. (2025). *23 Email Spam Statistics to Know in 2025.* https://www.mailmodo.com/guides/email-spam-statistics/

Kaggle, (2025). *The Enron Email Dataset.* https://www.kaggle.com/datasets/mohinurabdurahimova/maildataset

Khanday A., Shahbaz P., & Suraiya P. (2021). *Logistic Regression Based Classification of Spam and Non-Spam Emails.* https://doi.org/10.4108/eai.27-2-2020.2303291.

Mohammed, N., Mouhajir, M., & Yassine S.. (2023). *High Performance Computing Applied to Logistic Regression: A CPU and GPU Implementation Comparison.* https://www.researchgate.net/publication/373263539_High_Performance_Computing_Applied_to_Logistic_Regression_A_CPU_and_GPU_Implementation_Comparison

Papageorgiou, G., Economou, P., & Bersimis, S. (2024). A method for optimizing text preprocessing and text classification using multiple cycles of learning with an application on shipbrokers emails. *Journal of Applied Statistics*, 51(13), 2592–2626. https://doi: 10.1080/02664763./2024.2307535. PMID: 39290353; PMCID: PMC11404377.

Rakhmanov, O. (2020). A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments. *Procedia Computer Science,* 178, 194–204. https://doi.org/10.1016/j.procs.2020.11.021

Spam Statistics 2025 (2025). *New Data on Junk Email, AI Scams & Phishing.* https://www.emailtooltester.com/en/blog/spam-statistics/

Zhang, L., Ray, H., Priestley, J., & Tan, S. (2019). A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data. *Journal of Applied Statistics*, 47(3), 568–581. https://doi.org/10.1080/02664763.2019.1643829

References

Jyothiikaa Moorthy. (2025). *23 Email Spam Statistics to Know in 2025.* https://www.mailmodo.com/guides/email-spam-statistics/

Kaggle, (2025). *The Enron Email Dataset.* https://www.kaggle.com/datasets/mohinurabdurahimova/maildataset

Khanday A., Shahbaz P., & Suraiya P. (2021). *Logistic Regression Based Classification of Spam and Non-Spam Emails.* https://doi.org/10.4108/eai.27-2-2020.2303291.

Mohammed, N., Mouhajir, M., & Yassine S.. (2023). *High Performance Computing Applied to Logistic Regression: A CPU and GPU Implementation Comparison.* https://www.researchgate.net/publication/373263539_High_Performance_Computing_Applied_to_Logistic_Regression_A_CPU_and_GPU_Implementation_Comparison

Papageorgiou, G., Economou, P., & Bersimis, S. (2024). A method for optimizing text preprocessing and text classification using multiple cycles of learning with an application on shipbrokers emails. *Journal of Applied Statistics*, 51(13), 2592–2626. https://doi: 10.1080/02664763.2024.2307535. PMID: 39290353; PMCID: PMC11404377.

Rakhmanov, O. (2020). A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments. *Procedia Computer Science,* 178, 194–204. https://doi.org/10.1016/j.procs.2020.11.021

Spam Statistics 2025 (2025). *New Data on Junk Email, AI Scams & Phishing.* https://www.emailtooltester.com/en/blog/spam-statistics/

Zhang, L., Ray, H., Priestley, J., & Tan, S. (2019). A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data. *Journal of Applied Statistics*, 47(3), 568–581. https://doi.org/10.1080/02664763.2019.1643829

Отримано редакцією журналу / Received: 07.05.25
Прорецензовано / Revised: 16.05.25
Схвалено до друку / Accepted: 30.05.25

Viktor MOROZOV, PhD (Engin.), Prof.
ORCID ID: 0000-0001-7946-0832
e-mail: viktor.morozov@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Vladyslav DEINEHA, PhD Student
ORCID ID: 0009-0008-3123-2302
e-mail: thedynkan@knu.ua
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

SPAM DETECTION IN TEXT MESSAGES USING LOGISTIC REGRESSION BASED ON GRADIENT DESCENT

Background. ***With the increasing volume of email communication, the problem of spam filtering is becoming more and more relevant. According to statistical research, spam constitutes a significant portion of global email traffic, creating risks both for security and for the efficiency of electronic communication. In this context, natural language processing (NLP) and machine learning methods are gaining particular importance. The aim of this study is to develop a model for classifying email messages into spam and non-spam using logistic regression implemented via gradient descent, in combination with text data processing methods.***

Methods. ***The model was trained on a dataset containing over 5,000 email messages labeled as spam or non-spam. The data were preprocessed by removing noise components such as punctuation, numbers, stop words, and short tokens, followed by lemmatization. The cleaned texts were converted into numerical format using TF-IDF vectorization with L2 normalization. To address the class imbalance, the SMOTE method was applied. The model was trained using a classical gradient descent scheme with a sigmoid activation function and a logarithmic loss function.***

Results. ***The resulting model achieved high performance on the test set: overall accuracy was 98%, with an F1-score of 0.92 for the spam class and 0.99 for the non-spam class. The recall for spam reached 0.90, indicating the model's ability to detect most unwanted messages without excessive false positives. The balance between precision and recall is also reflected in macro and weighted average F1-scores, both exceeding 0.96.***

Conclusions. ***The findings demonstrate the effectiveness of combining logistic regression, gradient descent, and text preprocessing for the spam classification task, even in the presence of imbalanced data. The proposed approach is both efficient and interpretable, making it suitable for practical implementation in email filtering systems.***

Keywords: ***machine learning, natural language processing, gradient optimization, email filtering, text preprocessing, classification.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.

ISSN 2707-1758