МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ХАРЧОВИХ ТЕХНОЛОГІЙ

Інститут (факультет) Автоматизації і комп'ютерних систем
Кафедра Інформаційних технологій, штучного інтелекту і кібербезпеки

«До захисту в ЕК»
Директор інституту(декан факультету)
(підпис) Андрій ФОРСІЮК (ім'я та прізвище)
«13» грудня 2024р.

«До захисту допущено»
Завідувач кафедри
(підпис) Сергій ГРИБКОВ (ім'я та прізвище)
«13» грудня 2024р.

КВАЛІФІКАЦІЙНА РОБОТА
НА ЗДОБУТТЯ ОСВІТНЬОГО СТУПЕНЯ МАГІСТРА

зі спеціальності 122 «Комп'ютерні науки»
(код та назва спеціальності)
освітньо-професійної програми Управління інформацією та аналітика даних
на тему: Дослідження та розробка аналітичної системи з оцінки ефективності фільтрації спаму при кібератаках

Виконав: здобувач 2 курсу, групи КН-2-4М
Ковальчук Микола Васильович (підпис)
(прізвище, ім'я, по батькові повністю)

Керівник Чаплінський Юрій Петрович (підпис)
(прізвище , ім'я та по батькові повністю)

Консультанти (ім'я та прізвище) (підпис)
(ім'я та прізвище) (підпис)

Рецензент (ім'я та прізвище) (підпис)

Я як здобувач Національного університету харчових технологій розумію і підтримую політику університету з академічної доброчесності. Я не надавав(-ла) і не одержував(-ла) недозволеної допомоги під час підготовки цієї роботи. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

Здобувач (підпис)

Київ - 2024р.

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ХАРЧОВИХ ТЕХНОЛОГІЙ

Інститут (факультет) Автоматизації і комп'ютерних систем
Кафедра Інформаційних технологій, штучного інтелекту і кібербезпеки
Освітній ступінь магістр
Спеціальність 122 «Комп'ютерні науки»
(код і назва)
Освітньо-професійна програма Управління інформацією і аналітика даних
(назва)

ЗАТВЕРДЖУЮ
Завідувач
кафедри Інформаційних технологій, штучного інтелекту і кібербезпеки
Грибков С.В.
"07" **жовтня** **2024року**

З А В Д А Н Н Я

НА КВАЛІФІКАЦІЙНУ РОБОТУ ЗДОБУВАЧА

Ковальчук Микола Васильович
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження та розробка аналітичної системи з оцінки ефективності фільтрації спаму при кібератаках
керівник роботи Чаплінський Юрій Петрович доцент, канд.техн.наук
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом закладу вищої освіти від 7 жовтня 2024 року №884-кс
2. Строк подання здобувачем роботи 6 грудня 2024 року
3. Вихідні дані до роботи
Сучасні публікації зарубіжних авторів про спам.
Аналітичні та статистичні матеріали стосовно оцінки ефективності фільтрації спаму
Відомості про платформу Kaggle.
Технічна література з мови програмування Python
4. Зміст пояснювальної записки (перелік питань, які потрібно розробити)
Розділ 1.Аналіз предметної галузі та постановка задачі
Розділ 2. Дослідження ефективності фільтрації спаму при кібератаках
Розділ 3. Результати розробки аналітичної системи з оцінки ефективності фільтрації спаму при кібератаках
5. Перелік графічного матеріалу:
Види спаму та способи поширення; Структура аналітичної системи; Скріншоти результатів роботи аналітичної системи.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
1	доц.,канд.техн.наук. Чаплінський Ю.П.	07.10.2024	10.10.2024
2	доц.,канд.техн.наук. Чаплінський Ю.П.	17.10.2024	20.10.2024
3	доц.,канд.техн.наук. Чаплінський Ю.П.	27.10.2024	01.11.2024

7. Дата видачі завдання 7 жовтня 2024 року

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів виконання кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	Дослідження та аналіз предметної галузі	07.10.2024- 17.10.2024	Виконано
2	Дослідження ефективності фільтрації спаму при кібератаках	17.10.2024-27.10.2024	Виконано
3	Розробка аналітичної системи з оцінки ефективності фільтрації спаму при кібератаках	27.10.2024-17.11.2024	Виконано
4	Тестування системи з оцінки ефективності фільтрації спаму при кібератаках	17.11.2024-27.11.2024	Виконано
5	Дослідження отриманих результатів	28.11.2024-05.12.2024	Виконано
6	Оформлення кваліфікаційної роботи	07.10.2024-06.12.2024	Виконано
7	Оформлення автореферату	06.12.2024-16.12.2024	Виконано
8	Оформлення презентації	01.12.2024-06.12.2024	Виконано

Здобувач _____ (підпис) Ковальчук М.В. (прізвище та ініціали)

Керівник роботи _____ (підпис) Чаплінський Ю.П. (прізвище та ініціали)

АНОТАЦІЯ

Метою кваліфікаційної роботи є дослідження та розробка аналітичної системи з оцінки ефективності фільтрації спаму в умовах кібератак. В умовах постійного зростання обсягів спаму, що використовується для поширення шкідливого програмного забезпечення, фішингових атак та дезінформації, важливість вдосконалення методів кіберзахисту є надзвичайно актуальною.

У роботі проведено аналіз сучасних методів виявлення та фільтрації спаму, включаючи підходи на основі ключових слів, ймовірнісні алгоритми, машинне навчання та поведінкові моделі. Визначено їхні сильні та слабкі сторони, що дозволило сформулювати рекомендації для розробки аналітичної системи.

Розроблено функціонал системи фільтрації спаму та запропоновано ключові показники для її оцінки, такі як точність, повнота, специфічність, F1-міра та швидкість обробки даних. Проведено тестування системи на реальних наборах даних із платформи Kaggle, результати якого підтвердили її ефективність. Зокрема, досягнуто високих показників точності та продуктивності в умовах реальних загроз.

Розроблена система дозволяє ефективно оцінювати продуктивність різних методів фільтрації спаму, а також може бути інтегрована у сучасні системи кіберзахисту. Її впровадження сприятиме підвищенню рівня безпеки інформаційних систем підприємств і організацій.

Результати роботи мають практичне значення для подальшого вдосконалення методів фільтрації спаму та кіберзахисту і можуть бути використані як основа для розробки адаптивних систем захисту від кіберзагроз.

Структура та обсяг роботи. Пояснювальна записка курсової роботи складається з 3 розділів, містить 19 рисунків та 4 додатків.

Ключові слова: ФІЛЬТРАЦІЯ СПАМУ, КІБЕРАТАКИ, ІНФОРМАЦІЙНА БЕЗПЕКА, АНАЛІТИЧНА СИСТЕМА, ЗАХИСТ ДАНИХ, ШТУЧНИЙ ІНТЕЛЕКТ.

SUMMARY

The purpose of the master's work is research and development of an analytical system for evaluating the effectiveness of spam filtering in the context of cyber attacks. With the ever-increasing volume of spam used to spread malware, phishing attacks, and disinformation, the importance of improving cyber defense methods is extremely urgent.

The paper analyzes modern spam detection and filtering methods, including keyword-based approaches, probabilistic algorithms, machine learning, and behavioral models. Their strengths and weaknesses were determined, which made it possible to formulate recommendations for the development of an analytical system.

The functionality of the spam filtering system is developed and key indicators for its evaluation are proposed, such as accuracy, completeness, specificity, F1-measure, and data processing speed. The system was tested on real data sets from the Kaggle platform, the results of which confirmed its effectiveness. In particular, high levels of accuracy and productivity were achieved under conditions of real threats.

The developed system allows you to effectively evaluate the performance of various spam filtering methods, and can also be integrated into modern cyber protection systems. Its implementation will help increase the level of security of information systems of enterprises and organizations.

The results of the work are of practical importance for the further improvement of spam filtering and cyber protection methods and can be used as a basis for the development of adaptive systems of protection against cyber threats.Structure and scope of work. The explanatory note of the coursework consists of 3 sections, contains 19 figures.

Keywords: SPAM FILTERING, CYBER ATTACKS, INFORMATION SECURITY, ANALYTICAL SYSTEM, DATA PROTECTION, ARTIFICIAL INTELLIGENCE.

ЗМІСТ

Перелік скорочень, умовних позначень, термінів 8
ВСТУП 9
РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ 12
1.1. Актуальність роботи 12
1.2. Спам та антиспам 14
1.3. Класифікація спаму 21
1.4. Опис підходів щодо фільтрації спаму 29
1.5. Постановка задачі. 36
1.6. Висновок до першого розділу 37
РОЗДІЛ 2 ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ФІЛЬТРАЦІЇ СПАМУ ПРИ КІБЕРАТАКАХ 39
2.1. Процеси та правила фільтрації спаму 39
2.2. Сучасні методи фільтрації спаму 41
2.3. Підходи до оцінки ефективності фільтрації спаму 42
2.4. Показники ефективності фільтрації спаму 45
2.5. Висновки до другого розділу 50
РОЗДІЛ 3. РЕЗУЛЬТАТИ РОЗРОБКИ АНАЛІТИЧНОЇ СИСТЕМИ З ОЦІНКИ ЕФЕКТИВНОСТІ ФІЛЬТРАЦІЇ СПАМУ ПРИ КІБЕРАТАКАХ 52
3.1. Доцільність створення аналітичної системи з оцінки ефективності фільтрації спаму при кібератаках 52
3.2. Функціональні вимоги до системи з оцінки ефективності фільтрації спаму при кібератаках 54
3.3. Реалізація функцій аналітичної системи з оцінки ефективності фільтрації спаму при кібератаках 55
3.4 Результати праці аналітичної системи для оцінки ефективності фільтрації спаму при кібератаках 60
3.5. Висновки до третього розділу 68
ВИСНОВКИ 70

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 72
ДОДАТКИ 79
Додаток А. Код програми 79
Додаток Б. Скріншоти роботи програми 85
Додаток В. Схема основних етапів створення і розповсюдження спаму 88
Додаток Г. Представлення CSV файлу 89

Перелік скорочень, умовних позначень, термінів

TP - True Positives

TN - True Negatives

FP - False Positives

FN - False Negatives

ВСТУП

Актуальність теми. Розвиток сучасних технологій комунікацій, зокрема електронної пошти та соціальних мереж, створює середовище для поширення небажаних повідомлень, відомих як спам. Спам-повідомлення не тільки створюють інформаційний шум і знижують продуктивність користувачів, але й можуть становити реальну загрозу безпеці, поширюючи фішингові атаки, шкідливий код і дезінформацію. З огляду на це, розробка ефективних систем фільтрації спаму є нагальною задачею, яка вимагає адаптивних методів і сучасних технологій.

Мета дослідження. Метою кваліфікаційної роботи є дослідження та розробка аналітичної системи оцінки ефективності фільтрації спаму в умовах кібератак. Така система дозволить аналізувати методи фільтрації спаму, визначати їх сильні та слабкі сторони, а також адаптувати існуючі підходи до нових типів загроз.

Завдання дослідження. Для досягнення поставленої мети необхідно виконати такі завдання:

1. Провести аналіз сучасних методів виявлення та фільтрації спаму.
2. Визначити ключові показники для аналізу системи фільтрації спаму.
3. Реалізувати функціонал аналітичної системи
4. Провести її тестування на основі наборів даних, знайдений на платформі *Kaggle*.
5. Оцінити вплив розробленої системи на технічні показники безпеки інформаційних систем.

Об’єкт дослідження. Об’єктом дослідження є процеси та системи фільтрації спамових повідомлень у сучасних умовах кіберзагроз. У цих умовах спам став значною проблемою для інформаційних систем, зокрема у зв'язку з постійно зростаючими обсягами даних, а також різноманітністю типів спаму, таких як рекламні, фішингові та шкідливі повідомлення. Ці фактори ускладнюють традиційні методи фільтрації, що потребує нових підходів і адаптації фільтрів до швидко змінюваних технологічних умов.

Предмет дослідження. Предметом дослідження є аналітичні методи та алгоритми, які використовуються для виявлення, класифікації та фільтрації спаму в електронних повідомленнях.

Методи дослідження. У процесі дослідження використано такі методи:

- **Аналіз літературних джерел** для вивчення існуючих підходів до фільтрації спаму.
- **Методи машинного навчання**, включаючи ймовірнісні моделі та нейронні мережі, для аналізу та класифікації спамових повідомлень.
- **Емпіричне тестування** для оцінки ефективності аналітичної системи на наборі даних з *Kaggle*.
- **Порівняльний аналіз** для визначення переваг і недоліків різних методів фільтрації.

Наукова новизна одержаних результатів. Наукова новизна роботи полягає у розробці експериментальної аналітичної системи, що дозволяє ефективно оцінювати продуктивність алгоритмів фільтрації спаму. У роботі було інтегровано такі підходи:

- Розробка методики оцінки продуктивності, яка базується на ключових показниках ефективності, таких як точність, повнота, специфічність, F1-міра та швидкість обробки повідомлень.
- Застосування інструментів візуалізації, таких як ROC-криві та хмари слів, для полегшення аналізу результатів та виявлення закономірностей у характеристиках спамових повідомлень.
- Створення експериментальної платформи для тестування та порівняння ефективності різних алгоритмів фільтрації спаму на реальних наборах даних.

Практичне значення одержаних результатів. Практична цінність роботи полягає у створенні інструменту для аналізу ефективності методів фільтрації спаму. У рамках реалізації було використано набір інструментів *Python*,

включаючи бібліотеки *pandas*, *sklearn* та *seaborn*, що дозволяють не тільки обробляти великі обсяги даних, але й створювати зручні візуалізації для аналізу.

Особистий внесок здобувача. У рамках дослідження автором було:

- проведено аналіз літератури та сучасних методів фільтрації спаму;
- розроблено концепцію та структуру аналітичної системи;
- адаптовано набір даних з платформи *Kaggle* для потреб дослідження;
- реалізовано основні функції аналітичної системи;
- проведено тестування та інтерпретацію отриманих результатів.
- створено показники для подальшого аналізу.

Публікації. Результати дослідження апробовано на І Міжнародній науково-практичній конференції «Штучний інтелект та інформаційні технології» (АІІТ-2024), що відбулася 3–4 червня 2024 року у місті Київ, Україна. Результати представлені у вигляді тез доповідей:

Ковальчук М.В., Струзік В.А. Порівняльний аналіз популярних механізмів повнотекстового пошуку // «Штучний інтелект та інформаційні технології» (АІІТ-2024): наукові праці І Міжнародної науково-практичної конференції, м. Київ, Україна, 3–4 червня 2024 р. – К.: НУХТ, 2024. – С. 123–124.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1. Актуальність роботи

В сучасному цифровому світі інформаційна безпека стала невід'ємною складовою життєдіяльності як окремих людей, так і цілих організацій. Однією з основних загроз для інформаційних систем є спам — небажані повідомлення, які розсилаються масово та нерідко мають шкідливий або шахрайський характер. Спам використовується кіберзлочинцями як інструмент для різних видів кібератак, зокрема фішингу, поширення шкідливого програмного забезпечення, крадіжки даних, компрометації облікових записів та інших шкідливих дій. Таким чином, фільтрація спаму є важливою задачею, вирішення якої дозволяє суттєво знизити рівень кіберзагроз для користувачів та організацій.

За статистикою, щодня в інтернеті розсилаються мільярди небажаних електронних повідомлень, і ця цифра зростає з кожним роком. Поштові сервіси та корпоративні мережі стають першочерговими об'єктами атак, оскільки електронна пошта є основним засобом комунікації у бізнес-середовищі. Більше того, відповідно до останніх досліджень у галузі інформаційної безпеки, близько 85% усіх електронних повідомлень, що надсилаються щодня, є спамом. З огляду на обсяги таких даних, автоматизована обробка стає єдиним можливим способом виявлення та блокування спаму.

Актуальність дослідження проблеми фільтрації спаму зумовлена також стрімким розвитком методів та технологій, які використовуються кіберзлочинцями для обходу традиційних систем захисту. Традиційні фільтри спаму базуються на правилах, створених на основі досвіду роботи з попередніми повідомленнями, що призводить до ситуацій, коли вони швидко застарівають і не здатні розпізнати нові форми загроз. Зокрема, сучасні методи обходу спам-фільтрів включають використання випадкових символів у тексті повідомлення, шифрування вмісту, маскування *URL*-адрес та інші техніки, що роблять спам-повідомлення важко розпізнаваними для традиційних фільтрів.

Сьогодні ефективне фільтрування спаму вимагає використання сучасних алгоритмів машинного навчання та методів обробки природної мови. Машинне навчання дозволяє створювати адаптивні системи, які здатні аналізувати великі обсяги даних, враховувати різні характеристики повідомлень та швидко адаптуватися до нових видів загроз. Це робить такі алгоритми більш гнучкими та ефективними порівняно з традиційними методами на основі фіксованих правил. У зв'язку з цим, в останні роки популярності набули такі алгоритми, як наївний баєсів класифікатор, логістична регресія, методи градієнтного бустингу та глибокі нейронні мережі, які забезпечують високу точність класифікації спам-повідомлень.

Іншим аспектом актуальності даної роботи є економічні та соціальні наслідки, пов'язані з поширенням спаму. Спам-повідомлення, які потрапляють у корпоративні або особисті електронні поштові скриньки, не тільки створюють незручності для користувачів, але й призводять до значних фінансових витрат. Компанії змушені інвестувати значні ресурси в системи фільтрації, а також витрачати час та зусилля на видалення спаму та усунення його наслідків. Крім того, наявність спаму в електронних скриньках може призводити до втрати важливих повідомлень, компрометації корпоративної інформації та навіть витоку конфіденційних даних, що завдає репутаційної шкоди.

Зростання кількості та складності кібератак, а також використання спаму як одного з основних інструментів для реалізації таких атак, свідчить про нагальну потребу в дослідженні та вдосконаленні методів фільтрації спаму. Традиційні системи, що використовують фільтрацію на основі ключових слів чи списків заблокованих адрес, вже не здатні забезпечити належний рівень захисту. Саме тому сучасні наукові дослідження зосереджені на застосуванні більш складних моделей, здатних виявляти спам на основі аналізу патернів та поведінкових характеристик, що дозволяє значно підвищити точність класифікації.

Наукові інновації в галузі кібербезпеки та обробки природної мови також мають значний вплив на розвиток фільтрації спаму. Використання машинного навчання в поєднанні з алгоритмами аналізу тексту дозволяє створювати адаптивні системи, які здатні швидко реагувати на нові загрози та знижувати кількість

хибнопозитивних і хибнонегативних результатів. Такі підходи стають особливо важливими для забезпечення безпеки в умовах інтенсивної цифровізації суспільства, де дані стають основною цінністю і їх захист є пріоритетним завданням.

Таким чином, проблема фільтрації спаму є важливим напрямом у галузі інформаційної безпеки, і її актуальність обумовлена зростанням кількості кібератак, необхідністю захисту конфіденційної інформації та мінімізації економічних втрат, пов’язаних з поширенням небажаних повідомлень. Розробка та впровадження нових методів і моделей фільтрації спаму дозволить підвищити рівень захищеності інформаційних систем та сприятиме створенню безпечного цифрового середовища для користувачів та організацій.

1.2. Спам та антиспам

Поняття спаму

Спам (англ. *spam*) - Розсилка комерційної та іншої реклами або подібних комерційних видів повідомлень особам, які не висловлювали бажання їх отримувати [1].

Спам нерідко містить рекламну інформацію або ж використовується для фішингових та шкідливих атак. Спам-повідомлення зазвичай є низькоякісними та автоматично згенерованими, а їхнє розсилання спрямоване на отримання прибутку або реалізацію шкідливих намірів.

Поширення спаму створює значні проблеми для користувачів та адміністраторів мереж, адже через нього суттєво знижується ефективність роботи інформаційних систем, виникає навантаження на поштові сервери, а також збільшується ризик компрометації конфіденційних даних. Спам є серйозною загрозою для безпеки інформаційного середовища, оскільки повідомлення можуть містити посилання на фішингові сайти або ж вкладені файли зі шкідливим програмним забезпеченням (рис. 1.1).

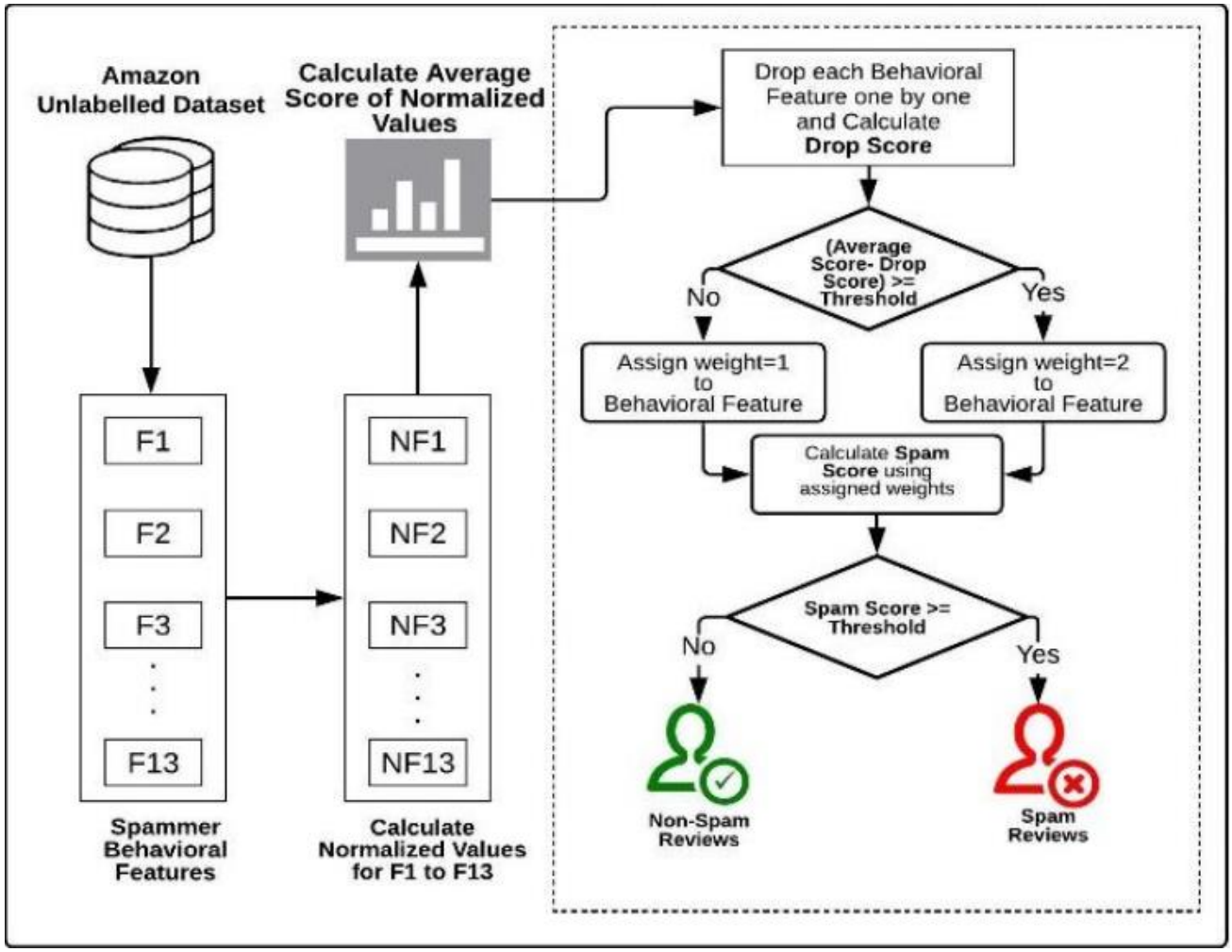

Рисунок 1.1 – Структура методу поведінки спамерів[2].

На рисунку вище виконуються чотири етапи:

1. Спочатку він обчислює нормалізоване значення (0-1) кожного спамера поведінкова особливість.

2. На основі цих значень обчислюється середній бал для кожного огляду та загальна точність і повний набір даних.

3. Далі оцінюється вплив кожного поведінкову функцію, дотримуючись методу скидання функції та призначає вагу відповідно до важливості кожного поведінкова особливість.

4. Нарешті, він обчислює показник спаму за допомогою зважені особливості поведінки та ідентифікує відгуки про спам і не спам, використовуючи різні порогові значення.

Основні типи спаму.

Спам можна поділити на кілька основних категорій (рис. 1.2):

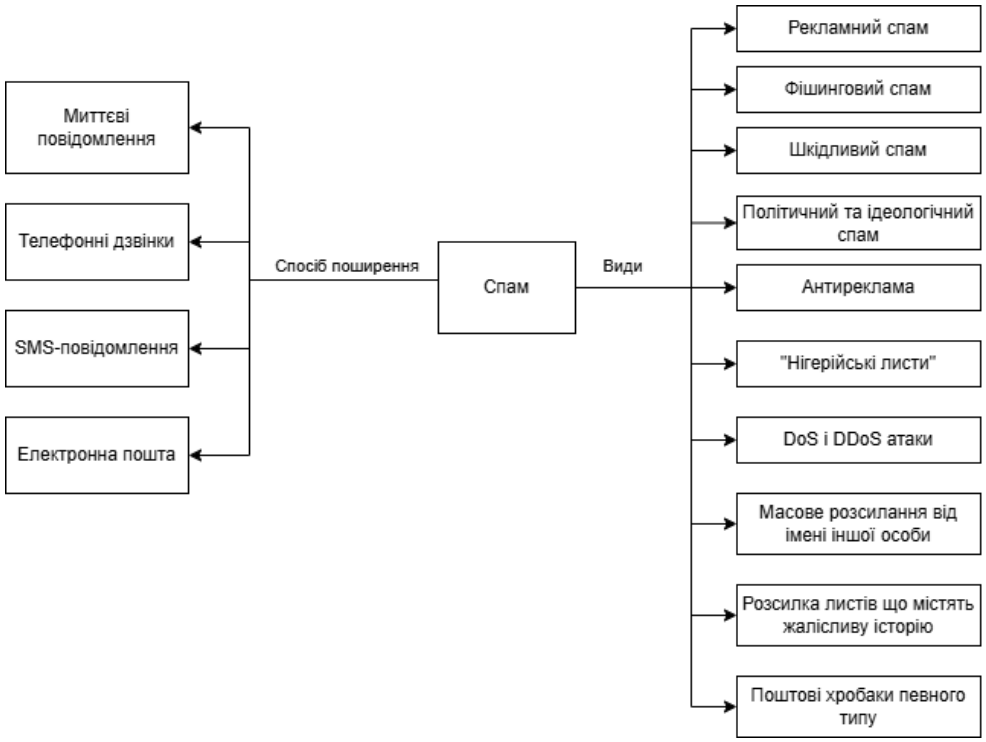

Рисунок 1.2 – Види спаму та способи поширення

Рекламний спам — повідомлення з комерційним змістом, націлені на просування товарів чи послуг [3].

Фішинговий спам — повідомлення, метою яких є викрадення конфіденційної інформації користувача, наприклад, логінів та паролів [4].

Шкідливий спам — повідомлення, які містять шкідливий код чи вкладені файли, призначені для зараження системи користувача.

Політичний та ідеологічний спам — повідомлення, які містять певні політичні чи соціальні ідеї з метою впливу на думку користувачів [5].

Антиреклама – це вид спаму, в якому використовуються образливі або провокаційні матеріали, спрямовані на підрив репутації компанії, особи чи бренду [5]. Зазвичай такі повідомлення можуть містити негативні відгуки або навмисно неправдиву інформацію, яка підриває довіру до суб’єкта.

«Нігерійські листи» - є одним з найстаріших типів шахрайського спаму. У таких листах відправники намагаються переконати одержувача надати особисту

інформацію чи фінансові дані під приводом виграшу, спадщини чи іншої винагороди. Популярність цих листів виникла ще в 90-х роках і досі продовжується з модифікаціями [6].

***DoS* і *DDoS*-атаки -** Спам використовується як інструмент для *DoS* (*Denial of Service*) та *DDoS* (*Distributed Denial of Service*) атак, коли масова розсилка завантажує сервери жертви, викликаючи збої в їх роботі. Зазвичай у таких атаках розсилається величезна кількість запитів, що перевантажують систему, роблячи її недоступною для звичайних користувачів [7].

Масове розсилання від імені іншої особи - Цей вид спаму полягає в тому, що шахраї маскуються під довірену особу (наприклад, колегу чи друга) та надсилають повідомлення з проханням надати доступ до конфіденційних даних або допомогти фінансово. Такі повідомлення часто виглядають як особисті, що підвищує ймовірність обману [8].

Розсилка листів, що містять жалісну історію - такий тип спаму часто пов’язаний із соціально інженерними атаками, де шахраї розповідають вигадану трагічну історію, щоб викликати співчуття й отримати фінансову допомогу від жертви. Це можуть бути історії про хвороби, нещасні випадки або втрату майна, які стимулюють людей переказати кошти [9].

Поштові хробаки певного типу - є шкідливими програмами, які поширюються через електронну пошту [10]. Зазвичай вони містять вкладення або посилання на шкідливий код, який активується під час відкриття. Після зараження комп'ютера хробаки автоматично надсилаються всім контактам з адресної книги користувача, що призводить до масового розповсюдження.

Основні ознаки спаму

1. Масовість та неперсоналізований характер

Спам, як правило, надсилається великим групам людей без їхнього дозволу. Такі повідомлення зазвичай мають загальний, неперсоналізований текст на кшталт "Шановний користувачу" або "Друже", замість імені одержувача. Відсутність персоналізації є ознакою того, що повідомлення може бути спамом.

2. Використання кричущих заголовків та емоційних закликів

Спамові повідомлення часто містять заголовки, що привертають увагу, наприклад, "Ви виграли!", "Сенсаційна пропозиція!", "Дійте негайно!" або "100% гарантія" [11]. Такі заклики створюють відчуття терміновості й спонукають до негайної реакції, не залишаючи часу на обдумування.

3. Наявність підозрілих або маловідомих посилань

Спам-повідомлення часто містять посилання на сторонні вебсайти, які можуть бути замасковані за допомогою скорочених або змінених *URL*-адрес [11]. Такі посилання можуть перенаправляти користувача на фішингові сайти або вебсайти, які поширюють шкідливе програмне забезпечення.

4. Пропозиція легких грошей або великих виграшів

Спам часто обіцяє швидке збагачення, виграш у лотереї, безкоштовні подарунки або інші вигоди [11]. Обіцянка великих фінансових винагород або "легких грошей" є однією з найпоширеніших ознак спаму, адже вона спрямована на спонукання до негайних дій з боку одержувача.

5. Граматичні помилки та нетиповий стиль написання

Спамери часто використовують автоматичні переклади або спрощені шаблони для створення своїх повідомлень, що призводить до граматичних і стилістичних помилок [11]. Незвичний стиль написання, помилки в тексті, а також дивні або нелогічні вислови можуть свідчити про спамовий характер повідомлення.

6. Підозрілі вкладення

Часто спам-повідомлення містять вкладення з підозрілими форматами файлів (наприклад, *.exe, .zip* або *.scr*), які можуть бути шкідливими [11]. Вкладення у вигляді документів із макросами або зашифрованих архівів також є поширеною ознакою шкідливого спаму.

7. Імітація відомих брендів або організацій

Спамери часто маскуються під відомі компанії або державні установи, використовуючи схожі логотипи, стилі оформлення або навіть домени, які візуально нагадують офіційні [11]. Однак доменні адреси можуть містити дрібні помилки або додаткові символи, що відрізняє їх від справжніх.

8. Використання занадто привабливих пропозицій і акцій

Спам-повідомлення часто містять акційні пропозиції з великими знижками, які виглядають занадто вигідними, щоб бути правдою [11]. Наприклад, пропозиція придбати новий *iPhone* за 10% від його ринкової вартості може свідчити про спам**.**

9. Наявність надлишкових або нав'язливих рекламних матеріалів

Спам-повідомлення часто містять численні банери, кнопки або візуальні елементи, які відволікають увагу. Надмірне використання графічних матеріалів та повторюваних закликів є ознакою того, що повідомлення може бути спамом.

Антиспам: методи боротьби зі спамом

Антиспам — це алгоритми та програмне забезпечення, метою яких є виявлення та блокування потенційно небезпечних електронних листів із скриньок вхідних повідомлень користувачів [12]. Протоколи захисту від спаму визначають, що таке небажане та небажане повідомлення (спам).

Боротьба зі спамом стала актуальним питанням для багатьох організацій та окремих користувачів. Для цього розроблено різноманітні антиспам-методи, які допомагають блокувати небажані повідомлення та знижувати ризик потрапляння спаму в особисту або корпоративну поштову скриньку.

Фільтрація за ключовими словами. Це один із найпростіших методів антиспаму, який передбачає пошук певних слів або фраз у тексті повідомлення [13]. Якщо виявлено заборонені слова або специфічні фрази, повідомлення позначається як спам. Проте цей метод має обмежену ефективність, оскільки кіберзлочинці можуть легко обходити його шляхом заміни символів, випадкових інтервалів або використання синонімів.

Використання чорних списків.Чорні списки містять адреси електронної пошти або *IP*-адреси, з яких раніше надходив спам [14]. Якщо відправник знаходиться у чорному списку, його повідомлення автоматично блокуються. Однак, через динамічні *IP*-адреси та можливість зміни адрес, цей метод також не завжди є надійним.

Аналіз заголовків повідомлень. Цей метод передбачає аналіз метаданих листів (зокрема, заголовків повідомлень) для виявлення підозрілих елементів, таких як відсутність даних про відправника або аномальні часи відправлення [14]. Даний підхід допомагає виявити спам на етапі доставки, але також може викликати проблеми, оскільки легітимні повідомлення можуть бути помилково позначені як спам.

Байєсовий фільтр.Метод, заснований на використанні наївного баєсового класифікатора, дозволяє фільтрувати спам шляхом аналізу частоти появи певних слів у спам-повідомленнях та звичайних повідомленнях [15]. Після навчання на обох типах листів, система може оцінювати ймовірність того, що нове повідомлення є спамом. Цей метод має високу ефективність і часто використовується в комбінації з іншими алгоритмами.

Машинне навчання та штучний інтелект. З розвитком технологій штучного інтелекту і машинного навчання з’явилися нові, більш ефективні методи фільтрації спаму. Зокрема, використання алгоритмів глибокого навчання дозволяє створювати моделі, здатні аналізувати текст повідомлення, структуру листа, а також виявляти спам на основі великої кількості параметрів [16]. Завдяки самонавчальній природі цих алгоритмів, вони можуть адаптуватися до нових загроз і ефективно працювати навіть за умов змінюваних типів спаму.

Антиспам-системи та їх компоненти. Ефективна антиспам-система зазвичай складається з кількох компонентів, що працюють спільно для досягнення максимальної ефективності в блокуванні спаму.

Основні компоненти таких систем включають:

База даних чорних списків — дозволяє швидко блокувати повідомлення від відомих джерел спаму.

Аналізатор тексту та метаданих — забезпечує перевірку наявності специфічних характеристик спаму, таких як відсутність заголовків або використання специфічних фраз.

Класифікатор — зазвичай реалізований на основі алгоритмів машинного навчання або байєсових фільтрів, призначений для визначення ймовірності, що повідомлення є спамом.

Блокувальник вкладених файлів — призначений для блокування потенційно небезпечних вкладень, таких як виконувані файли чи архіви з паролями.

Аналітичний модуль — використовується для моніторингу продуктивності антиспам-системи та її подальшого вдосконалення.

Виклики та недоліки сучасних антиспам-систем

Незважаючи на широкий арсенал методів та інструментів, сучасні антиспам-системи стикаються з численними викликами. Основними проблемами є високий відсоток помилково позитивних результатів, адаптивність спам-розсилок, а також зростання обсягів даних, які потребують обробки. Хибнопозитивні результати призводять до блокування легітимних повідомлень, що може завдати шкоди бізнес-процесам або особистим комунікаціям.

Адаптивність сучасних спам-розсилок є значним викликом для антиспам-систем. Зловмисники активно використовують методи уникнення, такі як вставка випадкових символів, приховання *URL*-адрес, зміна контексту, а також використання нових джерел розсилки. Це змушує антиспам-системи постійно оновлюватися та вдосконалюватися, аби протистояти новим типам загроз.

Зрештою, зростання обсягів інформації, яку потрібно обробляти, є ще однією важливою проблемою. Кількість електронної пошти, обмін повідомленнями та зростання мережевих ресурсів збільшують навантаження на антиспам-системи. Це вимагає не лише високої продуктивності фільтрів, а й оптимізації процесів обробки даних для зменшення затримок та збереження ефективності.

1.3. Класифікація спаму

Класифікація спаму є важливим аспектом для розуміння його різновидів і ефективного налаштування антиспам-систем. Різні типи спаму мають специфічні

характеристики, що впливають на методи виявлення та фільтрації. Нижче розглянуто основні категорії спаму та їх особливості.

1.3.1. Рекламний спам

Рекламний спам є найпоширенішим типом спаму (рис. 1.3), який часто має комерційну мету. Це повідомлення, що містять інформацію про товари, послуги або пропозиції, які користувачі не запитували [2]. Рекламний спам зазвичай надсилається масово на великі бази електронних адрес. Він може мати різну структуру та оформлення, проте головною його метою є залучення користувача до певних товарів чи послуг.

Характерними рисами рекламного спаму є:

- використання яскравих заголовків, щоб привернути увагу;
- посилання на веб-сайти чи сторінки товарів;
- візуальні елементи (зображення, банери), що підсилюють рекламне повідомлення.

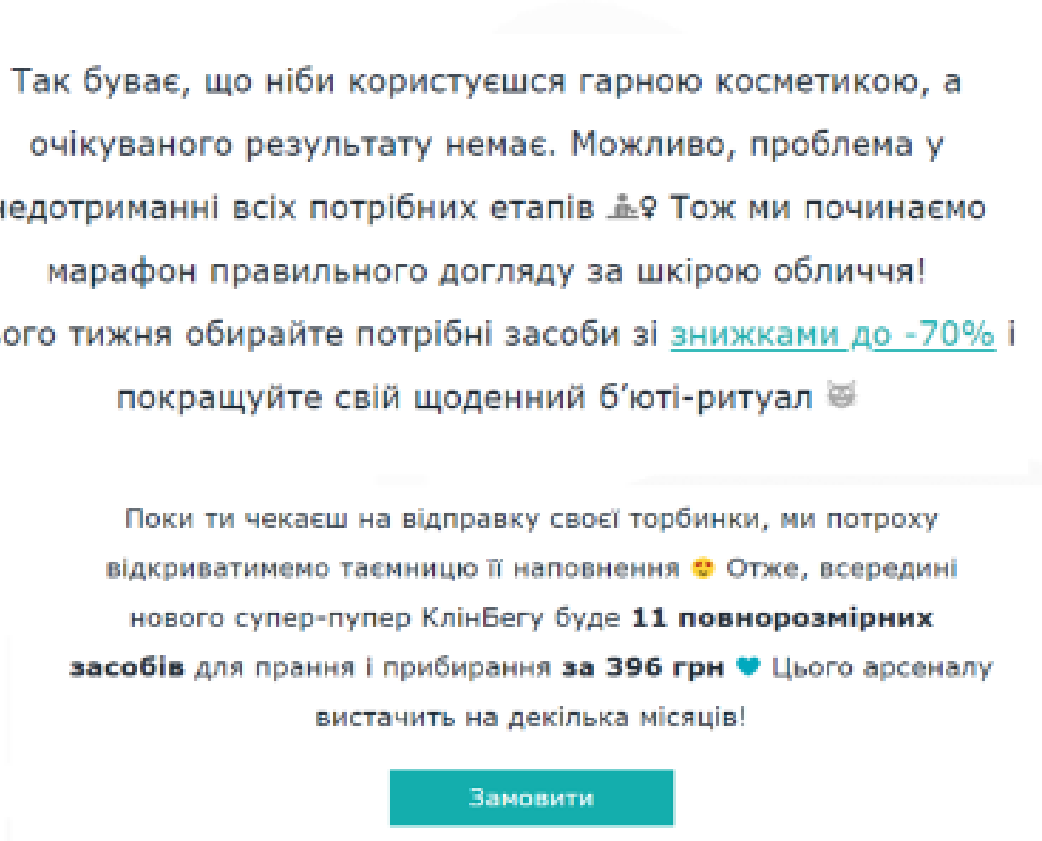

Рисунок 1.3 – Приклад спаму

У багатьох країнах рекламний спам заборонений на законодавчому рівні, оскільки його обсяг створює значне навантаження на інформаційні системи та поштові сервіси.

1.3.2. Фішинговий спам

Фішинговий спам — це один з найбільш небезпечних видів спаму, який націлений на викрадення конфіденційної інформації. Фішингові повідомлення часто імітують листи від офіційних установ, таких як банки, платіжні системи, державні установи або соціальні мережі [3]. Головною метою фішингового спаму є отримання доступу до особистих даних користувачів, таких як логіни, паролі, номери банківських карток або фінансова інформація.

Особливості фішингового спаму:

- маскування під офіційні листи (логотипи, оформлення, схожі домени);
- заклики до термінових дій (наприклад, "оновіть ваші дані" або "ваш акаунт заблоковано");
- шкідливі посилання, що перенаправляють на фальшиві веб-сайти, де збирається інформація (рис. 1.4).

Оскільки фішинговий спам представляє серйозну загрозу для користувачів, його виявлення є пріоритетом для багатьох антиспам-систем, що використовують додаткові заходи захисту, такі як аналіз поведінки *URL*-адрес і перевірка справжності відправника.

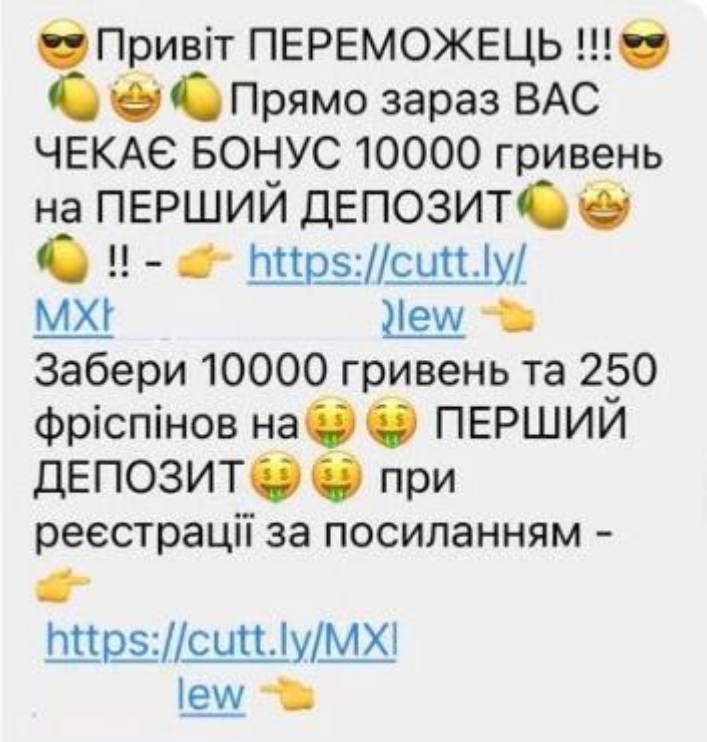

Рисунок 1.4. – Приклад фішингового спаму

1.3.3. Шкідливий спам (маліційний спам)

Шкідливий спам — це тип спаму, що містить шкідливий код або посилання на завантаження шкідливого програмного забезпечення [18]. Відкриття таких листів або завантаження вкладень може призвести до зараження комп'ютера або мережі користувача вірусами (рис.1.5), троянами або програмами-вимагачами (*ransomware*). Маліційний спам часто надсилається масово з метою зараження якнайбільшої кількості пристроїв.

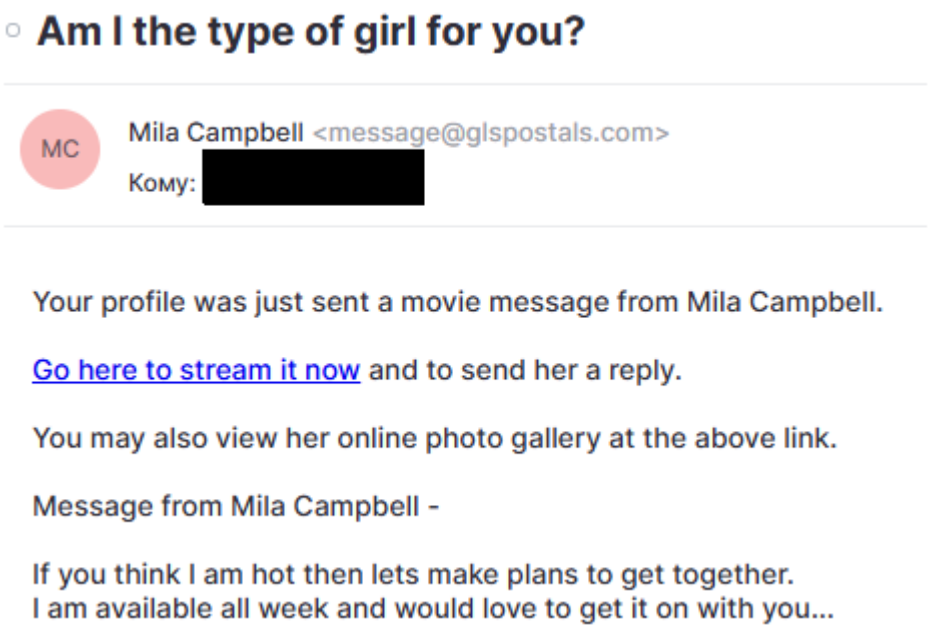

Рисунок 1.5 – Приклад шкідливого спаму

Основні риси шкідливого спаму:

- наявність вкладених файлів (наприклад, архівів, документів з макросами або виконуваних файлів);
- використання соціальної інженерії для стимулювання відкриття вкладення або завантаження файлу (наприклад, листи про "важливий рахунок" чи "документ з банку");
- посилання на заражені веб-сайти, які автоматично інфікують пристрій користувача.

Шкідливий спам становить загрозу як для індивідуальних користувачів, так і для організацій, тому його виявлення часто базується на поєднанні методів фільтрації за вмістом та поведінкових алгоритмів.

1.3.4. Соціально спрямований спам

Соціально спрямований спам зазвичай має на меті маніпулювання громадською думкою або просування певних ідей [19]. Він часто включає політичну, ідеологічну чи соціальну інформацію, а також може бути використаний

для дезінформації або пропаганди. Такий тип спаму може надходити через електронну пошту, соціальні мережі або коментарі на веб-сайтах.

Основні характеристики соціально спрямованого спаму:

- повідомлення часто мають сильний емоційний вплив;
- включають заклики до певних дій (наприклад, підписати петицію, приєднатися до акції);
- інформація може бути частково або повністю неправдивою (дезінформація).

Соціально спрямований спам є особливо небезпечним через можливий вплив на громадську думку, а також здатність дестабілізувати суспільні настрої. Для його виявлення застосовуються методи аналізу тональності тексту та визначення аномальних патернів поведінки в мережі.

1.3.5. Спам у соціальних мережах та месенджерах

Зі зростанням популярності соціальних мереж та месенджерів спам-повідомлення все частіше надсилаються через ці канали комунікації [19]. Такі повідомлення можуть містити рекламу, фішингові посилання або ж бути частиною шкідливих кампаній. Спам у соціальних мережах використовує не лише текстові повідомлення, але й зображення, відео та інші мультимедійні форми.

Характеристики спаму у соціальних мережах:

- використання ботів для автоматичного розповсюдження спаму;
- включення посилань на сторонні ресурси;
- візуальні елементи, що підвищують привабливість контенту (зображення, *GIF*).

Антиспам-засоби для соціальних мереж мають особливі вимоги, оскільки вони повинні враховувати специфіку платформи та швидкість розповсюдження інформації.

1.3.6. Інформаційний спам (інформаційний шум)

Інформаційний спам відрізняється від традиційного спаму тим, що він не завжди містить небажану рекламу або шкідливий код [20]. Це інформація, яка перевантажує користувача та відволікає його від основних завдань. До такого типу спаму можуть входити непотрібні повідомлення про оновлення, численні нагадування, повідомлення від мобільних додатків тощо.

Характеристики інформаційного спаму:

- повідомлення мають обмежену корисність для користувача;
- перевантаження інформацією може призводити до зниження продуктивності;
- не мають прямої загрози, але спричиняють когнітивне навантаження.

Інформаційний спам може бути небезпечним у великих обсягах, оскільки викликає так звану "інформаційну втому" та знижує здатність користувачів обробляти інформацію.

1.3.7. Підходи до класифікації спаму

Для класифікації спаму використовуються різні підходи, які базуються на аналізі вмісту, поведінкових характеристиках та структурі повідомлень. Основними підходами є:

Аналіз тексту — дозволяє класифікувати спам на основі ключових слів, тональності, шаблонів речень.

Аналіз поведінки відправника — відстежує шаблони розсилок, час відправлення та інші параметри.

Аналіз вкладених файлів і посилань — перевірка наявності шкідливих елементів у листах, як-от зловмисного коду або *URL*-адрес.

Машинне навчання та штучний інтелект — використання адаптивних алгоритмів, що навчаються розрізняти різні типи спаму на основі великої кількості ознак.

1.3.8. Технологічні особливості розповсюдження спаму

Основні етапи створення і розповсюдження спаму.

Створення та розповсюдження спаму можна поділити на кілька основних етапів. Спершу спамер розробляє саме повідомлення, яке може бути рекламним, фішинговим або шкідливим [21]. Це повідомлення зазвичай виглядає привабливо, щоб привернути увагу одержувача, з метою збору особистих даних, розповсюдження шкідливих програм чи просто для збільшення трафіку на певний сайт. Спамери також можуть використовувати техніки персоналізації, зокрема, звертання до користувача по імені або вказівку на його інтереси, що дозволяє зробити повідомлення більш переконливим.

Наступний етап — вибір методів розповсюдження. Зазвичай спамери використовують електронну пошту, оскільки це найбільш ефективний спосіб досягти великої аудиторії [22]. Вони можуть купувати списки електронних адрес або використовувати інструменти для збору адрес автоматично. Однак електронна пошта не є єдиним методом. Спамери також активно використовують соціальні мережі, месенджери та мобільні *SMS*, щоб охопити різні групи користувачів [21]. Розповсюдження через месенджери та соціальні мережі часто здійснюється через автоматизовані акаунти (боти), які відправляють повідомлення або публікують спам.

Згодом, щоб забезпечити успіх розсилки, спамер може вдатися до компрометації різних серверів [22]. Наприклад, спам може бути відправлений через захоплені сервери або за допомогою бот-мереж. Використовуються також слабкості в системах безпеки для того, щоб уникнути виявлення або блокування. Паралельно з технічними методами спамери можуть використовувати соціальні маніпуляції, такі як фішинг, щоб виманити дані користувачів або перенаправити їх на фальшиві вебсайти.

Крім того, важливим етапом є вдосконалення методів. Спамери ретельно моніторять свої кампанії, щоб зрозуміти, які з методів працюють найбільш ефективно [21]. Вони можуть коригувати свої стратегії, міняти домени,

використовувати нові підходи для обходу спам-фільтрів або змінювати вигляд повідомлень, щоб зробити їх менш підозрілими для автоматичних систем захисту. Це дозволяє їм адаптуватися до нових технологій, які з'являються для боротьби з розповсюдженням спаму.

Для протидії спаму існують різноманітні заходи. Найпоширенішими є фільтри спаму, які автоматично визначають і блокують непотрібні повідомлення. Крім того, організації та компанії проводять аналіз і звітність, щоб виявити нові методи спаму і вчасно реагувати на них. Законодавства деяких країн також передбачають штрафи за відправку небажаних повідомлень без дозволу користувачів, що є додатковим стримуючим фактором для спамерів.Схема основних етапів створення і розповсюдження спаму зображено в **додатку В.**

1.4. Опис підходів щодо фільтрації спаму

Фільтрація спаму є важливою складовою для захисту інформаційних систем від небажаних повідомлень та кіберзагроз. Різноманітність спаму вимагає застосування різних методів фільтрації, які можуть поєднуватися для досягнення найкращого результату.

1.4.1. Метод фільтрації на основі ключових слів

Фільтрація на основі ключових слів є одним із найпоширеніших підходів до блокування спаму. Вона полягає у використанні спеціальних словників, що містять ключові слова, які часто трапляються в спам-повідомленнях [13]. Де:

- $\{w_1, w_2, \dots, w_n\}$ — множина всіх слів, які використовуються для аналізу вхідного повідомлення.
- $\{k_1, k_2, \dots, k_m\}$ — множина ключових слів, які визначені як можливі індикатори спаму.
- M — це множина слів, що містяться у вхідному повідомленні.
- K — це множина ключових слів, які використовуються для аналізу і визначення, чи є повідомлення спамом.

$$M = \{w_1, w_2, \dots, w_n\}, K = \{k_1, k_2, \dots, k_m\} \quad (1.1)$$

Наприклад, слова типу "безкоштовно", "виграш", "клікніть тут" або "гроші" можуть вказувати на небажане повідомлення.

Переваги цього методу:

- Простота впровадження та низькі обчислювальні витрати.
- Ефективний для простого рекламного спаму.

Недоліки:

- Високий рівень помилкових спрацьовувань (*false positives*), оскільки слова з чорного списку можуть траплятися в легітимних повідомленнях.
- Спамери легко обходять фільтри, змінюючи написання слів (наприклад, "г-р-о-ш-і" замість "гроші").

1.4.2. Байєсівські фільтри

Байєсівські фільтри є популярним методом для фільтрації спаму, що базується на теоремі Байєса [15]. Цей підхід використовує ймовірнісний аналіз, щоб визначити, чи є повідомлення спамом на основі частотного аналізу слів у ньому. Спершу фільтр "навчається" на базі даних спамових та легітимних повідомлень, визначаючи ймовірність того, що конкретне слово належить до певного типу повідомлення. Після цього він може класифікувати нові повідомлення на основі накопиченого досвіду. Де:

- $P(Spam|M)$ — ймовірність того, що повідомлення є спамом.
- $P(M|Spam)P(M|\backslash text\{Spam\})P(M|Spam)$ — ймовірність спостереження конкретного вмісту повідомлення, якщо це спам.
- $P(Spam)P(\backslash text\{Spam\})P(Spam)$ — апостеріорна ймовірність спаму.
- $P(M)P(M)P(M)$ — загальна ймовірність такого повідомлення.

$$P(Spam|M) = \frac{P(M|Spam)*P(Spam)}{P(M)} \quad (1.2)$$

Переваги байєсівського методу:

- Ефективне самооновлення і здатність адаптуватися до нових видів спаму.
- Низький рівень помилкових спрацьовувань при достатньому обсязі навчальної вибірки.

Недоліки:

- Високі обчислювальні витрати для тренування моделі.
- Потребує достатньої кількості спамових та легітимних повідомлень для ефективного навчання.

1.4.3. Чорні списки та білі списки

Чорні списки та білі списки є одним із найпростіших підходів до фільтрації спаму. Чорний список містить *IP*-адреси або домени, з яких часто надходить спам, тоді як білий список включає адреси, яким можна завжди довіряти [14]. Якщо повідомлення надходить з *IP*-адреси або домену, що перебуває в чорному списку, воно автоматично блокується.

Переваги цього підходу:

- Висока ефективність при виявленні відомих джерел спаму.
- Простота реалізації та мінімальні обчислювальні витрати.

Недоліки:

- Неможливість блокування нових або випадкових джерел спаму.
- Ризик помилкового блокування легітимних повідомлень у разі динамічних *IP*-адрес.

1.4.4. Методи на основі машинного навчання

З розвитком технологій машинного навчання дедалі більше систем антиспам-фільтрації використовують адаптивні алгоритми, такі як дерева рішень, нейронні мережі та методи глибокого навчання. Ці алгоритми можуть розпізнавати патерни

у спамових повідомленнях на основі великої кількості ознак, таких як структура повідомлення, частотний аналіз слів, аналіз вмісту тощо.Де:

- $f(x)$ — функція класифікації, яка визначає, чи є повідомлення x спамом.
- $P(Spam|x)$ — ймовірність того, що повідомлення x є спамом, на основі його змісту.
- $Treshhold$ — граничне значення, що використовується для прийняття рішення. Якщо ймовірність перевищує цей поріг, повідомлення класифікується як спам; інакше — як "не спам".

$$f(x) = \begin{cases} Spam, \text{якщо } P(Spam|x) > Treshhold \\ Not\ Spam, \text{інакше} \end{cases} \quad (1.3)$$

Переваги методів машинного навчання:

- Висока точність і здатність до адаптації.
- Можливість обробки великої кількості характеристик, що підвищує ефективність виявлення складних спамових атак.

Недоліки:

- Високі витрати на обчислювальні ресурси, особливо для глибоких моделей.
- Необхідність у великих обсягах даних для тренування, що може бути обмеженням для деяких систем.

1.4.5. Метод аналізу поведінки (Behavioral Analysis)

Метод аналізу поведінки базується на спостереженні за характером та моделями відправлення повідомлень. Наприклад, якщо *IP*-адреса надсилає велику кількість повідомлень за короткий проміжок часу, це може вказувати на спам-активність. Аналіз поведінки є корисним для виявлення масових спам-кампаній та забезпечує оперативне реагування на нові загрози.

Переваги аналізу поведінки:

- Ефективний для блокування масових розсилок.
- Забезпечує захист від спаму, що не виявляється традиційними методами.

Недоліки:

- Може пропустити менш інтенсивні, але небезпечні види спаму.
- Ризик блокування легітимних листів при неправильній інтерпретації поведінки відправника.

1.4.6. Фільтрація за допомогою хешування (Checksum-based Filtering)

Хешування є ще одним підходом для фільтрації спаму, який використовується переважно для виявлення дублікатів спамових повідомлень. Цей метод передбачає створення контрольних сум (хешів) для повідомлень та порівняння їх із базою даних хешів спаму. Якщо хеш нового повідомлення збігається з хешем у базі даних, повідомлення класифікується як спам.

Переваги методу хешування:

- Ефективний для швидкого виявлення дублікатів спам-повідомлень.
- Низькі обчислювальні витрати на етапі порівняння.

Недоліки:

- Спамери можуть змінювати вміст повідомлень, щоб уникнути збігу хешів.
- Метод неефективний для унікальних повідомлень або спаму з випадковими модифікаціями.

1.4.7. Методи антиспуфінгу

1) Аутентифікація за допомогою цифрових підписів.

Цей метод використовує криптографічні технології для забезпечення автентичності повідомлень. Кожен відправник має унікальний цифровий підпис, який підтверджує, що повідомлення дійсно надійшло від зазначеного відправника. Для перевірки автентичності використовуються відкриті ключі, які дозволяють виявити підроблені повідомлення.

***S/MIME* (*Secure/Multipurpose Internet Mail Extensions*)** та ***PGP* (*Pretty Good Privacy*)** — популярні стандарти для підпису та шифрування електронної пошти, які забезпечують захист від підробки електронних листів.

***DKIM* (*DomainKeys Identified Mail*)** — використовує цифрові підписи для перевірки того, чи дійсно лист надійшов з вказаного домену.

2) Фільтрація за *IP*-адресою та методи *DNS*.

Фільтрація за *IP*-адресами — це один із основних методів захисту від спуфінгу в електронній пошті та веб-ресурсах. Метод полягає в перевірці *IP*-адреси відправника та порівнянні її з відомими списками.

***SPF* (*Sender Policy Framework*)** — це метод, за допомогою якого можна перевірити, чи може конкретна *IP*-адреса надсилати пошту від імені зазначеного домену.

***DMARC* (*Domain-based Message Authentication, Reporting, and Conformance*)** — покращує безпеку за допомогою перевірки як *SPF*, так і *DKIM*, і дозволяє вказати політику для обробки повідомлень, що не проходять перевірку.

3) Використання чорних та білих списків.

Це один із простих, але ефективних методів захисту від спуфінгу. Чорний список містить *IP*-адреси або домени, з яких надходять підозрілі або шкідливі повідомлення. Якщо повідомлення надходить з такої адреси, воно автоматично блокуються. Білий список включає довірені домени та *IP*-адреси, що дозволяє уникнути блокування легітимних повідомлень.

Чорний список — включає всі адреси та домени, з яких надходять спамові або фішингові повідомлення.

Білий список — гарантує, що повідомлення з цих джерел завжди будуть вважатися безпечними.

4) Машинне навчання для виявлення спуфінгу.

Системи машинного навчання здатні адаптуватися до нових типів атак та виявляти приховані патерни, які можуть свідчити про спуфінг *[*16*]*. Використовуються алгоритми класифікації, такі як дерева рішень, нейронні мережі

або методи глибокого навчання для аналізу великих обсягів даних та визначення наявності фальшивих елементів.

Нейронні мережі — використовуються для виявлення прихованих аномалій та нелінійних патернів в даних, що дозволяє ефективно виявляти нові типи спуфінгу.

Методи глибокого навчання — забезпечують виявлення складних патернів у великих обсягах даних, що дозволяє швидше реагувати на нові види атак.

5) Аналіз поведінки.

Аналіз поведінки базується на моніторингу і вивченні звичок користувачів або джерел повідомлень [23]. Зокрема, спостереження за аномаліями в поведінці відправників може допомогти виявити спуфінг.

Аналіз частоти надсилання повідомлень — якщо з певної *IP*-адреси надходить велика кількість повідомлень за короткий проміжок часу, це може свідчити про спам-активність.

Аналіз шаблонів поведінки — виявлення підозрілих патернів, таких як надсилання масових повідомлень з нових або змінних джерел.

6) Використання двофакторної аутентифікації (2*FA*).

Двофакторна аутентифікація забезпечує додатковий рівень захисту, який значно ускладнює підробку особи [24]. Навіть якщо зловмисник отримає доступ до пароля, без другого фактора (наприклад, одноразового коду або біометричних даних) доступ до системи неможливий.

7) Блокування спуфінгу через *DNS*.

Методи захисту на основі *DNS*, такі як ***DNSSEC*** (*DNS Security Extensions*) та ***DANE*** (*DNS-based Authentication of Named Entities*), дозволяють захистити домени від спуфінгу та фальшування сертифікатів [25].

DNSSEC — це набір розширень, що додає криптографічні підписи до записів *DNS*, запобігаючи атакам, спрямованим на зміну *DNS*-записів.

DANE — використовує *DNS* для підтвердження автентичності *TLS*-сертифікатів, що захищає від атак, спрямованих на фальшування веб-сайтів.

1.5. Постановка задачі.

Мета дослідження. Метою кваліфікаційної роботи є дослідження та розробка аналітичної системи оцінки ефективності фільтрації спаму в умовах кібератак.. Система повинна дозволяти аналізувати методи фільтрації спаму, визначати їх сильні та слабкі сторони, адаптувати існуючі підходи до нових типів загроз та оцінювати їх ефективність у реальних умовах кібератак.

Завдання дослідження. Проведений аналіз сучасних видів спаму та існуючих методів фільтрації спаму показав актуальність та необхідність проведення досліджень щодо оцінки ефективність фільтрації в умовах реальних загроз. Для цього необхідно

1. Провести аналіз сучасних методів виявлення та фільтрації спаму з метою дослідити переваги та недоліки існуючих підходів, таких як методи на основі ключових слів, байєсівська фільтрація, використання чорних списків, алгоритми машинного навчання та поведінкові моделі.Це дозволить виділити ключові переваги та недоліки кожного методу. Такі дослідження виконані в розділі

2. Визначити ключові показники для аналізу системи фільтрації спаму з метою подальшої розробки модель системи фільтрації спаму, яка б могла інтегрувати декілька методів одночасно для підвищення точності, включаючи машинне навчання для адаптації до нових загроз. Це базується на визначенні показників для оцінки ефективності системи на основі вибору метрик, таких як точність, повнота, швидкість обробки та частота помилкових спрацьовувань.

3. Визначити та реалізувати функціонал аналітичної системи. Для цього необхідно визначити функціонал аналітичної системи, який дозволив би проводити попередню обробку текстових даних, візуалізацію результатів та підтримку процесу аналізу отриманих результатів.

4. Провести тестування функціонування аналітичної системи на основі наборів даних платформи Kaggle. Це завдання передбачає використання даних реальних наборах даних із платформи Kaggle, що включають різні типи спаму (рекламні, фішингові, шкідливі повідомлення). Результати тестування дозволять

оцінити ефективність запропонованої системи та розробити рекомендації для зниження помилкових спрацьовувань.

5. Оцінити вплив розробленої системи на технічні показники безпеки інформаційних систем. Це дасть можливість виявити причини частих хибних позитивів та розробити рекомендації для зниження цьогопоказника шляхом удосконалення алгоритмів

1.6. Висновок до першого розділу

У першому розділі були розглянуті основні аспекти проблеми спаму, визначення його видів, а також сучасні методи фільтрації спаму. Окрім того, було проведено дослідження ключових типів спаму, таких як рекламний, фішинговий, шкідливий, соціально спрямований, а також спам у соціальних мережах і месенджерах. Також розглянуті основні підходи до фільтрації спаму, включаючи методи на основі ключових слів, байєсівську фільтрацію, використання чорних списків та машинне навчання.

Також детально розглянуті основні методи фільтрації спаму, зокрема, фільтрація за ключовими словами, байєсівські методи, чорні списки, а також використання сучасних підходів машинного навчання для виявлення спаму. Описано їхні переваги та недоліки.

Проведено класифікацію спаму та аналіз різних видів спаму, таких як рекламний, фішинговий, шкідливий, соціально спрямований, та спам у соціальних мережах, що допомагає у розробці адаптованих фільтрів для кожного з типів.

У цьому розділі також було розглянуто сучасні підходи до боротьби з новими видами спаму, зокрема застосування технологій машинного навчання, що дозволяють виявляти приховані патерни в повідомленнях і адаптувати фільтрацію до нових загроз.

Завдяки виконаним завданням це дозволяє сформувати чітке розуміння основних типів спаму та підходів до його фільтрації. Це є підґрунтям для подальшого розвитку аналітичної системи, яка буде виявляти і ефективно блокувати спам, адаптуючись до нових видів загроз.

Окремо були розглянуті такі підходи, як фільтрація за ключовими словами, байєсівські алгоритми, використання методів машинного навчання та аналіз мережевого трафіку. Незважаючи на досягнуті успіхи, усі ці методи мають свої обмеження, зокрема проблему помилкових спрацьовувань і недостатню адаптивність до нових типів спаму. У результаті виникла потреба у комбінуванні різних підходів для підвищення точності фільтрації.

Також було виконано оцінку методів фільтрації та їхніх недоліків створює основу для вдосконалення системи фільтрації, що дозволяє знизити кількість хибних спрацьовувань і підвищити ефективність виявлення спаму.

Можна зробити висновок про необхідність розробки більш гнучкої системи, яка б інтегрувала кілька підходів, використовуючи новітні технології, такі як штучний інтелект і глибинне навчання. Це стало підґрунтям для формування методології та реалізації аналітичної системи, опис якої буде детально представлений у другому розділі роботи.

Результати розділу показали необхідність проведення досліджень з аналізу ефективності методів фільтрації спаму

РОЗДІЛ 2 ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ФІЛЬТРАЦІЇ СПАМУ ПРИ КІБЕРАТАКАХ

2.1. Процеси та правила фільтрації спаму

Процес фільтрації спаму є багатоступеневим і складається з кількох етапів, які дозволяють виявляти, класифікувати та блокувати небажані повідомлення. Завдяки різноманітним підходам, таким як аналіз ключових слів, імовірнісний аналіз та машинне навчання, можна підвищити ефективність фільтрації спаму, знижуючи кількість помилкових спрацьовувань і адаптуючись до нових видів загроз. Розглянемо основні етапи та правила, які застосовуються в сучасних системах для боротьби зі спамом.

2.1.1. Основні етапи фільтрації спаму

Попередня обробка повідомлень:

- На цьому етапі відбувається збір і підготовка вхідних даних для подальшого аналізу [27]. Зокрема, система вилучає з повідомлень інформацію, яка може бути використана для ідентифікації спаму, як-от заголовок, відправник, *IP*-адреса, дата й час надсилання.
- Виконується базова обробка тексту: очищення від *HTML*-коду, видалення зайвих пробілів, транслітерація та перетворення всіх символів у нижній регістр для стандартизації даних [28].

Аналіз на основі правил:

- Використовуються заздалегідь визначені правила для відсіювання підозрілих повідомлень (формула 1.1) [29]. Наприклад, якщо відправник входить до чорного списку або повідомлення містить певні слова ("виграш", "гроші", "клікніть тут"), воно може автоматично позначатися як спам.
- Перевага цього етапу в його швидкості та простоті, однак такий метод має високий ризик помилкових спрацьовувань, оскільки шаблонний підхід не завжди дозволяє точно ідентифікувати спам.

Аналіз контенту повідомлень:

- Сучасні системи застосовують аналіз текстового контенту для виявлення ознак спаму. Це може бути перевірка наявності частих ключових слів, аналіз структури фраз і стилю написання.
- Алгоритми використовують статистичні методи для оцінки частоти ключових слів або специфічних словосполучень, які найчастіше трапляються в спамових повідомленнях.

Імовірнісний аналіз (напр. Байєсівська фільтрація):

- Алгоритм обчислює ймовірність того (формула 1.2), що повідомлення є спамом, на основі частотного аналізу певних слів або фраз. Такий підхід є гнучким і дозволяє адаптуватися до нових видів спаму.
- Імовірнісний аналіз вимагає навчання системи на наборі даних, які містять як спамові, так і легітимні повідомлення. Це дозволяє алгоритму ефективно виявляти спам і знижувати кількість помилкових спрацьовувань.

Аналіз на основі поведінкових моделей:

- Цей підхід включає моніторинг поведінки користувачів та відправників, що дозволяє виявляти масові розсилки або нетипову активність, яка може вказувати на спам.
- Наприклад, якщо з певної *IP*-адреси надходить велика кількість повідомлень за короткий період, це може свідчити про розсилку спаму. Також враховуються показники, як-от частота відкриття листів або кліків на посилання, що дозволяє оцінити достовірність повідомлення.

Машинне навчання та штучний інтелект:

- Використання алгоритмів машинного навчання дозволяє створювати адаптивні системи фільтрації (формула 1.3), які навчаються на великих обсягах даних. Моделі, як-от нейронні мережі або дерева рішень, можуть виявляти складні закономірності у спамових повідомленнях, підвищуючи ефективність фільтрації.
- Навчання таких моделей вимагає значних обчислювальних ресурсів і великої вибірки навчальних даних, але вони забезпечують високу точність і адаптивність, що важливо для захисту від нових кіберзагроз.

2.1.2. Основні правила фільтрації спаму

У сучасних системах фільтрації спаму використовуються наступні основні правила:

- **чорний список**: автоматичне блокування повідомлень, що надходять із джерел, внесених до чорного списку. До нього можуть входити певні *IP*-адреси, домени або навіть адреси електронної пошти, які раніше ідентифікувалися як джерела спаму [14].
- **білий список**: дозволяє пропускати повідомлення від довірених джерел без додаткової перевірки. Білий список використовується для запобігання помилкових блокувань листів від перевірених відправників [30].
- **частотне обмеження**: обмежує кількість повідомлень, які можуть бути надіслані з однієї *IP*-адреси або домену за певний проміжок часу. Якщо перевищено ліміт, подальші повідомлення можуть бути заблоковані або перевірені додатково [31].
- **перевірка автентичності**: включає методи, як-от *SPF* (*Sender Policy Framework*) та *DKIM (DomainKeys Identified Mail*), які підтверджують, що повідомлення надіслано з довіреного джерела. Це знижує ризик фішингових атак і зловмисного спаму [32].
- **аналіз посилань**: перевірка всіх гіперпосилань, що містяться в повідомленні, на предмет їхньої безпеки. Якщо посилання веде на підозрілий або шкідливий сайт, повідомлення може бути позначено як спам [33].
- **системи оцінювання**: базуються на сукупності різних критеріїв (наприклад, кількість ключових слів, результати імовірнісного аналізу, наявність підозрілих посилань). Система надає повідомленню підсумковий бал, і якщо він перевищує певний поріг, повідомлення класифікується як спам.

2.2. Сучасні методи фільтрації спаму

Сучасні системи фільтрації спаму базуються на поєднанні традиційних підходів та інноваційних технологій, таких як штучний інтелект і машинне

навчання. Це дозволяє виявляти не лише очевидні спам-повідомлення, але й складні загрози, що імітують легітимну комунікацію.

Методи на основі машинного навчання використовують алгоритми, які аналізують широкий спектр характеристик повідомлень — від тексту та метаданих до стилістичних особливостей. Завдяки здатності до самонавчання такі методи забезпечують високу точність і адаптивність. Однак їх впровадження вимагає значних обчислювальних ресурсів і великих обсягів даних для навчання.

Семантичний аналіз і методи обробки природної мови (NLP) дозволяють аналізувати контекст та зміст повідомлень. Ці підходи особливо ефективні у виявленні спаму, який приховано імітує легітимні повідомлення. Проте їх використання потребує складних алгоритмів і значних обчислювальних витрат.

Гібридні системи інтегрують кілька підходів, зокрема поведінковий аналіз, машинне навчання та сигнатурні методи. Такі системи характеризуються високою надійністю та універсальністю, хоча і мають складність у реалізації та високу залежність від ресурсів.

Антиспуфінгові технології, такі як SPF, DKIM і DMARC, зосереджені на перевірці автентичності відправника через налаштування доменних політик і криптографічні підписи. Це підвищує рівень безпеки електронної пошти, але не завжди ефективно захищає від складних маніпуляцій.

Інноваційним напрямом є використання блокчейн-технологій у системах антиспаму. Децентралізовані платформи забезпечують перевірку репутації відправників і збереження історії взаємодій. Хоча блокчейн пропонує прозорість і стійкість до атак, його впровадження ускладнюється високими вимогами до ресурсів і складністю інтеграції.

2.3. Підходи до оцінки ефективності фільтрації спаму

Оцінка ефективності фільтрації спаму є важливою складовою розробки та вдосконалення систем, що займаються захистом від небажаних повідомлень. Система фільтрації спаму повинна бути здатна точно і швидко ідентифікувати шкідливі чи небажані повідомлення, одночасно мінімізуючи ймовірність

помилкових спрацьовувань, коли легітимні повідомлення потрапляють у категорію спаму. Для цього існує кілька підходів, що базуються на різних принципах оцінки, зокрема, кількісних, які дозволяють виміряти точність фільтрації, і якісних, що оцінюють досвід користувачів та загальну ефективність системи.

Один із найбільш поширених способів оцінки ефективності фільтрації спаму – це застосування математичних метрик, таких як точність, чутливість, специфічність та інші. Вони дозволяють дати чітке уявлення про якість роботи фільтраційної системи. Точність (*accuracy*) є однією з найбільш вживаних метрик. Вона визначає, яку частину всіх повідомлень система класифікує правильно, тобто, як спам, так і нормальні повідомлення. Однак точність не завжди може бути надійним індикатором, оскільки вона не враховує важливість помилок класифікації. Наприклад, якщо система класифікує велику кількість спамових повідомлень, але при цьому помилково позначає важливі листи як спам, це може серйозно вплинути на якість фільтрації.

Чутливість (*recall*), або також відома як повнота, є важливим показником для оцінки здатності системи виявляти спам серед усіх спамових повідомлень. Вона відповідає на питання: «скільки відсотків від усіх спамових повідомлень система правильно класифікує як спам?». Цей показник важливий, оскільки система може мати високу точність, але водночас пропускати спамові повідомлення, що теж має важливе значення для загальної ефективності.

Не менш важливою є специфічність (*specificity*), яка вимірює здатність системи правильно відкидати нормальні повідомлення, не позначаючи їх як спам. Висока специфічність означає, що система здатна мінімізувати кількість помилкових спрацьовувань. Це особливо важливо для користувачів, оскільки помилкові спрацьовування можуть призвести до того, що важливі повідомлення не будуть отримані або не будуть помічені вчасно.

Однак лише точність, чутливість та специфічність не дають повної картини ефективності фільтрації спаму. Важливим аспектом є також швидкість обробки повідомлень, тобто час, який система витрачає на класифікацію кожного листа. Система повинна працювати не лише точно, але й швидко, оскільки у випадку

великих обсягів інформації затримки можуть стати суттєвою проблемою. Швидкість фільтрації особливо важлива для організацій, що обробляють великі обсяги електронної пошти, оскільки затримки можуть призвести до зниження ефективності всієї інфраструктури.

Для оцінки ефективності також важливо враховувати вплив фільтрації спаму на загальну продуктивність системи. Проблеми, що виникають при обробці великих масивів даних, можуть стати суттєвим бар'єром для реалізації ефективної фільтрації. Наприклад, необхідність обробляти великі обсяги інформації може викликати навантаження на сервери або знижувати ефективність роботи інших компонентів. Таким чином, під час оцінки фільтраційних систем необхідно враховувати не тільки їх здатність відрізняти спам від нормальних повідомлень, але й їх вплив на інші аспекти роботи інфраструктури.

Важливим аспектом є також тестування адаптивності до нових форм спаму. Оскільки спам постійно еволюціонує, система повинна бути здатна до постійного оновлення своїх правил і методів фільтрації. Оцінка ефективності має враховувати, як швидко система може адаптуватися до нових типів спаму, а також наскільки гнучко вона може інтегрувати нові моделі та алгоритми, що використовують машинне навчання чи інші сучасні методи аналізу даних. Залучення таких технологій дозволяє фільтраційним системам не лише виявляти відомі типи спаму, але й реагувати на нові загрози, яких не було в попередніх наборах даних.

Методи оцінки також можуть включати статистичний аналіз, який дозволяє виявити закономірності в поведінці спаму, зокрема його еволюцію з часом. Аналіз трендів допомагає виявляти нові форми атаки, що дозволяє системі швидше реагувати на зміни. Однак статистичні методи потребують великої кількості даних для навчання та тестування, що може бути проблемою у випадку обмеженості доступних даних або високої вартості збирання цих даних.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (2.1)$$

Формула оцінки середніх частот спамових атрибутів (формула 2.1).

Де:

- μ — середнє значення атрибута (наприклад, частота використання ключового слова).
- x_i — значення атрибута для i-го повідомлення.
- N — кількість повідомлень.

Наступною важливою властивістю є крос-валідація, яка дозволяє оцінити надійність фільтраційної системи, шляхом її перевірки на різних підмножинах даних. Це дозволяє уникнути проблеми переобучення, коли система добре працює тільки на певних даних, але не здатна адаптуватися до нових умов. Крос-валідація надає більш об'єктивну оцінку та допомагає забезпечити стійкість та універсальність фільтрації.

Фільтрація спаму в умовах постійно змінюваних кіберзагроз потребує особливої гнучкості та швидкості. Технології повинні бути здатні швидко адаптуватися до нових форм атак і зберігати високу точність навіть у випадках, коли зловмисники використовують більш складні методи, такі як соціальна інженерія чи автоматизовані бот-мережі. Система повинна забезпечувати не тільки високу ефективність у виявленні спаму, але й швидке реагування на нові загрози, що з'являються з часом. Оцінка її ефективності повинна включати не лише традиційні показники, такі як точність і чутливість, але й додаткові метрики, які враховують швидкість обробки повідомлень та здатність системи працювати в реальному часі, при цьому знижуючи ймовірність помилкових спрацьовувань. Це дозволить забезпечити надійний захист від спаму, навіть у складному та змінному технологічному середовищі.

2.4. Показники ефективності фільтрації спаму

Показники ефективності фільтрації спаму є ключовим інструментом для оцінки якості роботи систем, що займаються автоматичним відбором небажаних повідомлень в електронній пошті або інших платформах [34]. Вони дозволяють

визначити, наскільки система здатна правильно класифікувати повідомлення як спам чи не спам, а також оцінити її вплив на продуктивність та зручність для користувачів. Оскільки фільтрація спаму є важливою частиною багатьох інформаційних систем, правильне визначення та аналіз цих показників дозволяє розробникам, аналітикам та організаціям удосконалювати існуючі рішення та адаптувати їх до нових умов і викликів. Визначимо основні показники ефективності фільтрації спаму, а також роль кожного з них в оцінці системи.

2.4.1. Точність (Accuracy)

Точність — це найпоширеніший показник для оцінки ефективності фільтраційної системи (формула 2.2). Він визначає, яку частину всіх повідомлень система класифікує правильно, тобто і спамові повідомлення, і нормальні (не спамові) повідомлення.

- TP — кількість правильно класифікованих спамових повідомлень,
- TN — кількість правильно класифікованих не спамових повідомлень,
- FP — кількість нормальних повідомлень, які були помилково класифіковані як спам,
- FN — кількість спамових повідомлень, які були помилково класифіковані як не спам.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.2}$$

Точність є простим та загальним показником ефективності фільтрації, однак вона не дає повного уявлення про якість роботи системи. Наприклад, якщо система класифікує лише незначну кількість спаму, точність може залишатися високою, навіть якщо більша частина спамових повідомлень не була заблокована.

2.4.2. Чутливість (Recall)

Чутливість або повнота — це показник (формула 2.3) який вимірює здатність системи правильно визначати спам серед усіх спамових повідомлень. Він дає змогу зрозуміти, скільки з усіх спамів система змогла виявити.

$$Recall = \frac{TP}{TP+FN} \tag{2.3}$$

Чутливість є дуже важливим показником для систем фільтрації спаму, оскільки дозволяє оцінити, наскільки добре система справляється з завданням виявлення спамових повідомлень. Однак, як і точність, чутливість не є ідеальним показником, оскільки її високе значення може бути досягнуто за рахунок великої кількості помилкових спрацьовувань.

2.4.3. Специфічність (*Specificity*)

Специфічність — це показник, який вимірює здатність системи правильно відкидати нормальні (не спамові) повідомлення (формула 2.4), не класифікуючи їх як спам. Він важливий для того, щоб уникнути помилкових спрацьовувань, коли важливі повідомлення від користувачів потрапляють до спам-папки.

$$Specifity = \frac{TN}{TN+FP} \tag{2.4}$$

Висока специфічність означає, що система здатна мінімізувати кількість помилкових спрацьовувань. Для кінцевого користувача це важливо, оскільки мінімізація помилок дає змогу не втрачати важливі повідомлення, що знижує ймовірність недоречних збоїв у комунікації.

2.4.4. Помилкові спрацьовування (*False Positive Rate*)

Помилкове спрацьовування — це ситуація, коли система класифікує нормальне повідомлення як спам. Висока кількість помилкових спрацьовувань може призвести до того, що важливі повідомлення не будуть отримані вчасно. Помилкові спрацьовування — це важливий показник (формула 2.5), оскільки він має безпосередній вплив на зручність користування фільтром.

$$FPR = \frac{FP}{FP+TN} \tag{2.5}$$

Якщо система має високе значення помилкових спрацьовувань, то користувачі можуть втратити важливу інформацію, що зробить використання такої системи неприємним.

2.4.5. Точність позитивного прогнозу (*Precision*)

Точність позитивного прогнозу є показником, який визначає, скільки з усіх повідомлень, які система класифікувала як спам, насправді є спамом.

$$Precision = \frac{TP}{TP+FP} \tag{2.6}$$

Цей показник дозволяє оцінити, наскільки надійно система класифікує повідомлення як спам (формула 2.6). Висока точність позитивного прогнозу зменшує кількість помилкових спрацьовувань, однак часто для її досягнення доводиться жертвувати чутливістю.

2.4.6. *F*-міра

F-міра (або *F*1-*score*) є комбінованим показником (формула 2.7), який враховує як точність, так і чутливість.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2.7)$$

F-міра є важливим показником, оскільки вона дозволяє знайти компроміс між чутливістю і точністю. Це особливо корисно, коли є необхідність забезпечити баланс між кількістю виявлених спамів та мінімізацією помилкових спрацьовувань.

2.4.7. Час обробки

Час обробки повідомлення — це важливий показник, що вимірює швидкість, з якою система класифікує повідомлення. Оскільки фільтрація спаму часто відбувається в реальному часі, швидкість обробки є критично важливою для користувачів і систем. Час обробки залежить від складності алгоритмів, на яких побудована система, і може бути різним для різних типів фільтрації, таких як блеклістінг або аналіз вмісту повідомлень.

2.4.8. Показник ROC AUC (Area Under Curve)

Це показник (формула 2.8), який оцінює здатність класифікаційної моделі відрізняти між собою два класи, у нашому випадку — спам і не спам. *ROC* (*Receiver Operating Characteristic*) крива будується на основі співвідношення *True Positive Rate* (чутливість) та *False Positive Rate* (помилкові спрацьовування) при зміні порогу класифікації.

$$TPR = \frac{TP}{TP + FN} \quad (2.8)$$

AUC (*Area Under Curve*) — площа під цією кривою, яка варіюється від 0 до 1. Значення, близьке до 1, свідчить про високу ефективність моделі: вона точно відокремлює класи незалежно від вибору порогу. Якщо *AUC* становить 0.5, це вказує на випадкову класифікацію, а значення менше 0.5 означає, що модель працює гірше, ніж випадкове вгадування.

2.5. Висновки до другого розділу

У другому розділі були вирішені основні завдання, сформульовані в вступі, що стосуються дослідження ефективності фільтрації спаму в умовах кібератаки. Зокрема, було розглянуто процеси та методи фільтрації спаму, зокрема традиційні методи, такі як фільтрація на основі ключових слів і чорних списків, а також сучасні підходи, зокрема машинне навчання, байєсівська фільтрація, аналіз поведінки та гібридні системи. Оцінка ефективності фільтрації спаму здійснювалася через основні метрики, включаючи точність, чутливість, специфічність, *F*-міру, а також показник *ROC AUC*, що дозволяє оцінити здатність моделі відрізняти спам від не спаму.

Було виконано:

- Аналіз сучасних методів виявлення та фільтрації спаму — проведено детальний аналіз існуючих підходів до фільтрації спаму, таких як використання ключових слів, байєсівська фільтрація, чорні та білі списки, а також методи машинного навчання і поведінковий аналіз.
- Оцінка ефективності фільтрації спаму — оцінено основні показники, такі як точність, чутливість, специфічність та *AUC*, що дозволяють повноцінно оцінити ефективність фільтрації.
- Визначення основних показників для аналізу системи фільтрації спаму — розглянуті ключові метрики та їх роль у процесі фільтрації спаму, що дозволяє оцінити як точність класифікації, так і швидкість обробки даних.

Результати цього розділу підкреслюють важливість гнучкості та адаптивності фільтраційних систем, оскільки спам постійно еволюціонує. Для досягнення високої ефективності необхідно інтегрувати різні методи та стратегії, забезпечуючи точну і швидку фільтрацію спаму.

Результати, що отримані при виконанні розділу, показали необхідність розробки програмних засобів у вигляді аналітичної системи для проведення оцінки ефективності фільтрації спаму в реальних умовах кібератак, яка б дала змогу

проведення фахівцями всебічних досліджень з аналізу ефективності методів фільтрації спаму.

РОЗДІЛ 3. РЕЗУЛЬТАТИ РОЗРОБКИ АНАЛІТИЧНОЇ СИСТЕМИ З ОЦІНКИ ЕФЕКТИВНОСТІ ФІЛЬТРАЦІЇ СПАМУ ПРИ КІБЕРАТАКАХ

3.1. Доцільність створення аналітичної системи з оцінки ефективності фільтрації спаму при кібератаках

У сучасному цифровому середовищі спам є одним з найбільш поширених інструментів для проведення кібератак. Спам не тільки обмежує ефективність обробки електронної пошти, але й використовується для прихованих атак, таких як фішинг, розповсюдження вірусів та зловмисних програм. Оскільки спам є основним каналом для більшості кібератак, ефективна фільтрація спаму є критично важливою для забезпечення безпеки інформаційних систем. Тому створення аналітичної системи для оцінки ефективності фільтрації спаму є актуальним завданням для організацій, що прагнуть підвищити рівень захисту від кібератак.

Доцільність створення такої системи випливає з кількох важливих факторів. По-перше, постійне удосконалення методів фільтрації спаму дозволяє знизити кількість небажаних повідомлень, що потрапляють до системи, і таким чином зменшує кількість можливих шкідливих впливів. Однак, навіть найсучасніші фільтри не здатні забезпечити 100% захисту, що підвищує важливість створення системи, яка дозволяє оцінити ефективність фільтрації і виявляти слабкі місця у системі безпеки.

По-друге, з кожним роком методи атак стають все більш складними і адаптованими до нових умов, що вимагає постійного вдосконалення механізмів фільтрації. Аналітична система дозволяє виявити нові види спаму, визначити зони для поліпшення існуючих фільтрів і навіть передбачати можливі вектори нових атак.

Третім важливим аспектом є необхідність моніторингу ефективності фільтрації в реальному часі. Це дозволяє оперативно реагувати на нові загрози, оптимізувати систему фільтрації та приймати обґрунтовані рішення для удосконалення системи захисту. Система також дозволяє здійснювати аналіз

помилкових спрацьовувань, що важливо для забезпечення високої точності роботи фільтрації.

Підтверджуючи важливість таких систем, було проаналізовано в роботі «Порівняльний аналіз популярних механізмів повнотекстового пошуку» було розглянуто повнотекстові пошукові механізми, які здатні ефективно обробляти великі обсяги даних, застосовувати складні запити та виконувати масштабовану індексацію [36]. Аналогічно, ефективність аналітичних систем фільтрації спаму залежить від їх здатності до масштабування, адаптації до нових умов та високої продуктивності.

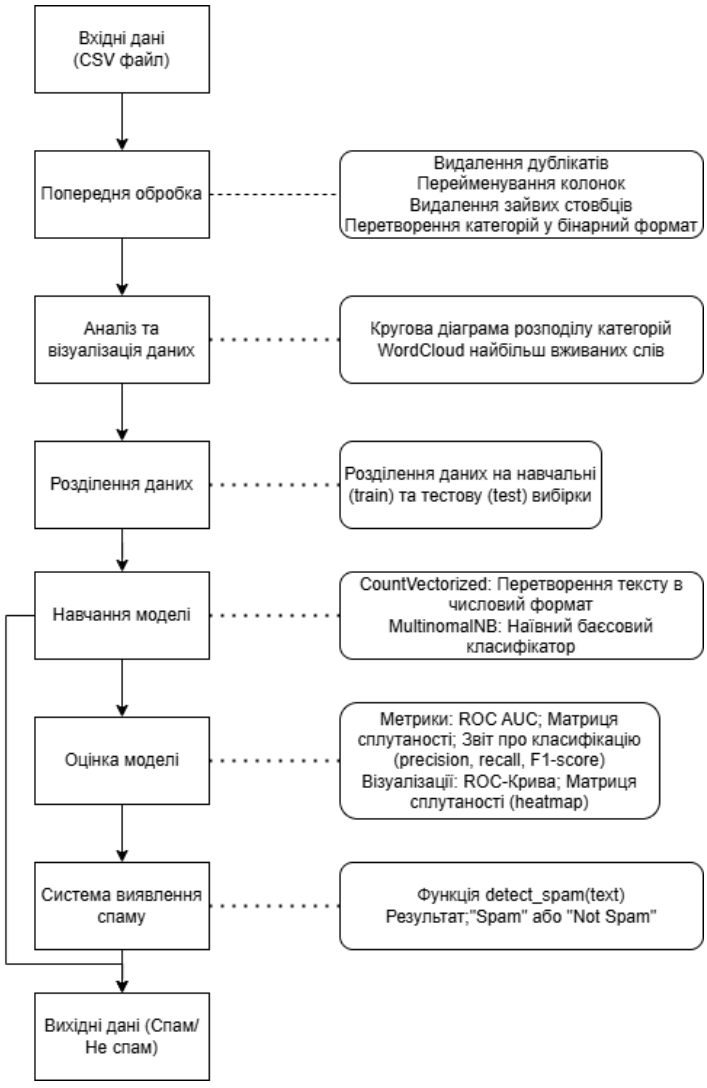

Рисунок 3.1 – Алгоритм роботи аналітичної системи

Загалом, створення аналітичної системи з оцінки ефективності фільтрації спаму є важливою складовою частиною стратегії кібербезпеки. Така система

дозволяє підвищити ефективність боротьби з кіберзагрозами, знижує ризик проникнення шкідливих програм та покращує захист персональних та корпоративних даних. Для кращого проектування системи було створено алгоритм роботи аналітичної системи (рис. 3.1).

3.2. Функціональні вимоги до системи з оцінки ефективності фільтрації спаму при кібератаках

Система для оцінки ефективності фільтрації спаму при кібератаках повинна забезпечувати наступні функціональні можливості:

Обробка текстових даних:

- Система повинна мати можливість отримувати, зберігати та обробляти текстові повідомлення для подальшого аналізу.
- Повідомлення повинні бути представлені у вигляді текстових рядків, що можуть містити як звичайний текст, так і *HTML* або вбудовані посилання.

Порівняння ефективності фільтрації:

Система повинна надавати порівняння між результатами фільтрації спаму для різних наборів даних, таких як:

- Оцінка ефективності алгоритмів фільтрації на основі метрик, таких як точність (*accuracy*), точність виявлення спаму (*precision*), повнота (*recall*), *F*1-міра, *ROC AUC*.
- Порівняння результатів на тренувальних та тестових даних для визначення стабільності моделі.

Візуалізація результатів.

Система повинна забезпечувати візуалізацію результатів аналізу через графіки та діаграми:

- Графік розподілу спаму та не спаму (*Pie chart*).
- ***WordCloud*** для візуалізації найбільш використовуваних слів у спам-повідомленнях.
- ***ROC*-крива** для порівняння результатів тренувальних та тестових даних.

- **Матриця плутанини (*Confusion Matrix*)** для візуалізації точності класифікації.
- **Звіт про класифікацію (*Classification Report*)** для більш детального аналізу результатів класифікації.

Оцінка моделей:

Система повинна підтримувати можливість оцінки різних моделей фільтрації спаму (наприклад, наївний баєс, логістична регресія, *SVM*) і вибір найбільш ефективної моделі на основі результатів оцінки.

Оцінка ефективності повинна проводитися на основі стандартних метрик, таких як:

- точність (*accuracy*);
- точність виявлення спаму (*precision*);
- повнота (*recall*);
- *F*1-міра;
- *ROC AUC;*
- час, витрачений на класифікацію.

Звітність і аналітика:

Після завершення аналізу система повинна генерувати звіт, який містить:

- порівняння результатів фільтрації;
- оцінку ефективності для кожного класу (спам/не спам);
- технічні деталі (наприклад, точність, *ROC AUC, F*1-міра тощо);
- рекомендації щодо покращення фільтрації спаму в майбутньому.

Завдяки такомим вимогам до функціоналу дозволяють створити систему, яка ефективно оцінює і покращує процес фільтрації спаму, що є важливим елементом при захисті від кіберзагроз.

3.3. Реалізація функцій аналітичної системи з оцінки ефективності фільтрації спаму при кібератаках

На етапі реалізації функцій аналітичної системи використовувались сучасні методи обробки даних та алгоритми машинного навчання, які є основою для

ефективної фільтрації спаму. Для навчання та тестування моделей було використано набір даних із платформи ***Kaggle***, яка є однією з найбільших платформ для роботи з відкритими наборами даних.

Kaggle — це платформа для аналітиків, дослідників даних і розробників машинного навчання, яка дозволяє ділитися наборами даних, проводити експерименти, брати участь у змаганнях і вдосконалювати навички в обробці даних та побудові моделей.

Однією з основних можливостей платформи є доступ до великої кількості різноманітних наборів даних, які можна використовувати для тренування моделей і тестування алгоритмів. *Kaggle* також надає інтерактивне середовище для програмування, де користувачі можуть писати код, запускати його і публікувати свої результати у вигляді "кернелів" (кернел — це код, який можна поділитися з іншими). Важливою частиною *Kaggle* є спільнота, де учасники можуть обмінюватися ідеями, задавати питання та допомагати один одному у вирішенні складних задач.

Kaggle служить потужним інструментом для розвитку навичок в аналізі даних і машинному навчанні, дозволяючи користувачам не лише покращувати свої здібності, а й знаходити практичні рішення для реальних задач, співпрацюючи з іншими дослідниками та отримуючи зворотний зв'язок.

Обраний набір даних містить різноманітну інформацію про текстові повідомлення, їхні ознаки (наприклад, частота вживання певних слів, довжина повідомлення тощо) та відповідну класифікацію як "спам" або "не спам". Ці дані були попередньо оброблені, щоб підготувати їх до моделювання.

1. Обробка текстових даних

Першим кроком було завантаження датасету із файлу *CSV*, що містить повідомлення. Ці дані можуть бути текстовими і містити два основні стовпці: "*Category*" (який позначає категорію повідомлення — спам чи не спам) і "*Message*" (який містить сам текст повідомлення).Цей код забезпечує підготовку вхідного датасету для подальшого аналізу та моделювання:

Завантаження даних:

df = pd.read_csv("*spam.csv*", *encoding='ISO*-8859-1')

Завантажуємо дані із *CSV*-файлу. Датасет містить повідомлення (як спам, так і звичайні) разом із додатковою інформацією.

Перевірка структури даних. Для початку визначається кількість рядків і стовпців, а також типи даних, що містяться в датасеті. Це допомагає зрозуміти структуру даних і визначити наявність пропущених значень, які потребують обробки.

print("Кількість рядків: ", *df.shape[0]*)

print("Кількість стовпців: ", *df.shape[1]*)

df.info()

Код вище використовується для отримання загальної інформації про кількість рядків, стовпців, типи даних та пропущені значення.

Попередня обробка:

df.rename(*columns*={"*v*1": "*Category*", "*v*2": "*Message*"}, *inplace=True*)

df.drop(*columns*={'*Unnamed*: 2', '*Unnamed*: 3', '*Unnamed*: 4'}, *inplace=True*)

df['Spam'] = df['Category'].apply(*lambda x*: 1 *if x* == '*spam*' *else* 0)

Перейменовуємо колонки для кращого розуміння. Змінено назви стовпців для зручності, наприклад, "*v1*" на "*Category*", а "*v2*" на "*Message*". Це покращує зрозумілість даних.

Видаляємо зайві стовпці, які не несуть корисної інформації. Оскільки деякі стовпці не містять корисної інформації (наприклад, "*Unnamed*: 2", "*Unnamed*: 3"), вони були видалені. Це зменшує розмір даних і спрощує подальшу обробку.

Додаємо бінарний стовпець *Spam*: 1 для спаму, 0 для не-спаму. Створено новий стовпець "*Spam*", в якому для кожного повідомлення позначено, чи є воно спамом (1) або не спамом (0). Це спрощує подальше машинне навчання, оскільки моделі краще працюють з числовими даними, ніж з текстовими.

2. Візуалізація

Цей код допомагає зрозуміти структуру даних за допомогою графіків.

Кругова діаграма. Використана для візуалізації розподілу спамових і не спамових повідомлень. Це дозволяє наочно побачити, який клас переважає в датасеті, що важливо для балансування моделі.

spread = df['Category'].value_counts()

spread.plot(*kind*='*pie*', *autopct*='%1.2*f*%%', *cmap*='*Set*1')

Відображає розподіл спам- і не-спам-повідомлень у відсотках.

Хмара слів. Створена для візуалізації найбільш вживаних слів у спам-повідомленнях. Це дає змогу зрозуміти, які терміни найчастіше зустрічаються в спамі, що допомагає в подальшому аналізі.

comment_words = ''

for val in df_spam.Message:

tokens = *str*(*val*).*split*()

comment_words += " ".*join*(*tokens*).*lower*() + " "

wordcloud = *WordCloud*(*width*=1000, *height*=500, *stopwords*=*STOPWORDS*).*generate*(*comment_words*)

plt.imshow(*wordcloud*)

Генерує візуалізацію найбільш вживаних слів у спам-повідомленнях. Це допомагає зрозуміти, які слова найчастіше зустрічаються.

3. Розробка моделі

На цьому етапі створюється і навчається модель для класифікації.

Поділ даних:

X_train, X_test, y_train, y_test = train_test_split(*df.Message*, *df.Spam*, *test_size*=0.25)

Розділяємо дані на тренувальну (75%) і тестову (25%) вибірки. Дані розділені на 75% для тренування моделі і 25% для тестування. Це дозволяє перевірити, наскільки добре модель працює на нових, не бачених даних.

Створення моделі:

clf = Pipeline([

('*vectorizer*', *CountVectorizer*()), # Текстова векторизація

('*nb*', *MultinomialNB*()) # Класифікація за допомогою Наївного Баєса

])

CountVectorizer: перетворює текст у числові вектори (мішок слів). Використовували метод *CountVectorizer*, який перетворює текстові повідомлення в числові вектори. Це дозволяє алгоритмам машинного навчання оперувати з текстовими даними у вигляді числових ознак, таких як частота використання слів.

MultinomialNB: модель класифікації для текстових даних. Використано для класифікації спамових та не спамових повідомлень. Алгоритм використовує ймовірнісні методи для передбачення, до якого класу належить повідомлення.

4. Оцінка моделі

Для визначення точності фільтрації використовуються метрики продуктивності:

Оцінка метрик:

roc_auc_train = roc_auc_score(*y_train, y_pred_train*)

accuracy_train = accuracy_score(*y_train, y_pred_train*)

*f*1*_train = f*1*_score*(*y_train, y_pred_train*)

Обчислюємо:

ROC AUC: показник здатності моделі розрізняти спам та не-спам. Оцінюється здатність моделі розрізняти спам і не спам. *ROC*-крива дозволяє побачити, як змінюється точність моделі в залежності від порогу класифікації. *AUC* (*Area Under Curve*) вимірює площу під цією кривою, де значення 1 вказує на ідеальну модель, а 0.5 — на випадкову класифікацію.

Accuracy: частка правильно класифікованих повідомлень.

F*1-*Score: баланс між точністю і повнотою. Оскільки точність і чутливість можуть бути суперечливими, *F*1-*Score* допомагає знайти компроміс.

Побудова *ROC*-кривої:

fpr_train, tpr_train, _ = roc_curve(*y_train, pred_prob_train*)

plt.plot(*fpr_train, tpr_train, label="Train ROC AUC*:
{:.2*f*}".*format*(*roc_auc_train*))

ROC-крива показує співвідношення між рівнем помилкових спрацювань та справжньою позитивною частотою.

Матриця сплутаності. Візуалізує кількість правильно та неправильно класифікованих спамових та не спамових повідомлень, що дозволяє оцінити, де модель допускає помилки.

cm_train = confusion_matrix(y_train, y_pred_train)

sns.heatmap(cm_train, annot=True, cmap="Oranges")

Відображає кількість правильних і хибних класифікацій.

5. Детекція спаму. Для класифікації нових повідомлень, система приймає текстове повідомлення, передає його в модель і повертає результат: чи є це спамом або ні.

Функція *detect_spam* приймає текст повідомлення і повертає результат:

def detect_spam(email_text):

prediction = clf.predict([email_text])

return "Це спам (*Spam*)!" *if prediction* == 1 *else* "Це не спам (*Ham*)!"

Текст обробляється в моделі.

Повертається результат у вигляді текстового повідомлення.

3.4 Результати праці аналітичної системи для оцінки ефективності фільтрації спаму при кібератаках

1. Опис результатів аналізу датасету

Під час обробки та аналізу вихідного датасету були отримані наступні результати (рис. 3.2):

```
Кількість рядків:  5572
Кількість стовпців:  5
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
Кількість дубльованих рядків: 403
```

Рисунок 3.2 – Виконаний код програми обробки даних

Структура даних:

- Кількість рядків у датасеті: **5572**.
- Датасет не містить пропущених значень після очищення.

Попередня обробка даних:

Видалено дублікати та непотрібні стовпці.(Видалення дублікатів покращує точність моделі, забезпечуючи чистоту та унікальність кожного запису в датасеті, а видалення непотрібних стовпців допомагає зменшити обсяг даних, позбавивши систему від зайвих і непотрібних елементів.)

Створено бінарний стовпець *Spam*, що значно спростило аналіз і підготовку даних до навчання (рис. 3.3).

```
[ ]  # Перевірка кількості унікальних значень для кожної змінної за допомогою циклу
     for i in df.columns.tolist():
         print("Кількість унікальних значень у стовпці", i, "становить", df[i].nunique())

     Кількість унікальних значень у стовпці v1 становить 2
     Кількість унікальних значень у стовпці v2 становить 5169
     Кількість унікальних значень у стовпці Unnamed: 2 становить 43
     Кількість унікальних значень у стовпці Unnamed: 3 становить 10
     Кількість унікальних значень у стовпці Unnamed: 4 становить 5

[ ]  # Зміна назв стовпців v1 та v2 на Category та Message
     df.rename(columns={"v1": "Category", "v2": "Message"}, inplace=True)

[ ]  # Видалення всіх стовпців без імен (які містять багато пропущених значень)
     df.drop(columns={'Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'}, inplace=True)

[ ]  # Створення бінарного стовпця 'Spam': 1 для 'spam' і 0 для 'ham', на основі стовпця 'Category'.
     df['Spam'] = df['Category'].apply(lambda x: 1 if x == 'spam' else 0)
```

Рисунок 3.3 – Виконаний код перевірки пропущених значень

Машинне навчання ефективніше працює з числовими даними, ніж з текстовими. Також використання текстових даних потребувало б додаткового попереднього оброблення для перетворення категорій на числові значення; Числове значення знижує ймовірність помилок у трактуванні категорій. З числовими даними зручно працювати під час створення графіків та діаграм.

2. Візуалізація результатів

Кругова діаграма розподілу спаму.

Рисунок 3.4 – Діаграма розроділу спаму

- Наочно продемонстровано, що більшість повідомлень у наборі є не-спам.
- На діаграмі (рис. 3.4) є 2 види розподілу повідомлень “*spam*” (*Spam* — це небажані або шкідливі повідомлення) та “*ham*” (це нормальні листи).
- Діаграма підкреслює нерівномірний розподіл, що може вплинути на балансування моделі.

Де:

- P_{ham} — ймовірність того, що повідомлення не є спамом (легітимне).
- N_{ham} — кількість легітимних (не спам) повідомлень у наборі даних.
- $N_{\text{загальні}}$ — загальна кількість повідомлень у наборі даних (включає і спам, і не спам).
- P_{spam} — ймовірність того, що повідомлення є спамом.
- N_{spam} — кількість спам-повідомлень у наборі даних.

Частка повідомлень: $P_{ham} = \frac{N_{ham}}{N_{\text{загальні}}}$; $P_{spam} = \frac{N_{spam}}{N_{\text{загальні}}}$. (3.1)

Якщо $N_{ham} = 4825, N_{spam} = 747$ тоді далі маємо такі обчислення:

$$P_{ham} = \frac{4825}{5572} \approx 0.866\ (86.6\%);\ P_{spam} = \frac{747}{5572} \approx 0.134\ (13.4\%) \quad (3.2)$$

Хмара слів для спаму:

- Найпоширеніші слова в спам-повідомленнях включають *"FREE," "call," "claim," "won," "mobile."*
- Ця візуалізація вказує на ключові слова (рисунок 3.5), які часто використовуються в спам-атаках.

Використовуються частоти слів у текстах. $f_{\text{слово}} = \frac{N_{\text{вживань слова}}}{N_{\text{загальні слова}}}$; (3.3)

Наприклад слово “*FREE*”: $f_{FREE} = \frac{100}{10000} = 0.01\ (1\%)$; (3.4)

Де:

- $f_{\text{слово}}$ — Частота вживання конкретного слова в тексті або наборі текстів.
- $N_{\text{вживань слова}}$ – Кількість разів, коли дане слово з'являється в тексті або наборі текстів.
- $N_{\text{загальні слова}}$ — Загальна кількість усіх слів у тексті або наборі текстів.

Рисунок 3.5 – Візуалізація найбільш вживаних сліва у спам повідомленнях

3. Ефективність моделі машинного навчання

Після навчання моделі та її тестування на вибірці, були отримані наступні метрики:

Точність (*Accuracy*):

$$A_{train} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{544+3608}{544+3608+10+17} = \frac{4152}{4179} \approx 99.35\% \quad (3.5)$$

$$A_{test} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{173+1205}{173+1205+2+13} = \frac{1378}{1393} \approx 98.92\% \quad (3.5)$$

- Тренувальна вибірка: 99.35%
- Тестова вибірка: 98.92%
- Високий показник точності підтверджує, що модель успішно класифікує більшість повідомлень.

➢ **Повнота (*Recall*)**:

$$R_{train} = \frac{TP}{TP+FN} = \frac{544}{544+17} = \frac{544}{561} \approx 0.9700 = 97\% \quad (3.6)$$

$$R_{test} = \frac{TP}{TP+FN} = \frac{173}{173+13} = \frac{173}{186} \approx 0.9301 = 93.01\% \quad (3.6)$$

- Для спам-повідомлень на тренувальній вибірці**:** 97%
- Для спам-повідомлень на тестовій вибірці**:** 93.01%

- Ця метрика показує, що модель здатна успішно розпізнавати більшість спамових повідомлень, що є ключовим для безпеки.

Презиційність (*Precision*):

Частка правильно класифікованих як спам повідомлень (істинно позитивних) серед усіх повідомлень, які модель позначила як спам (істинно позитивні + хибно позитивні).

$$P_{train} = \frac{TP}{TP+FP} = \frac{544}{544+10} = \frac{544}{554} \approx 0.9820 = 98.2\% \quad (3.7)$$

$$P_{test} = \frac{TP}{TP+FP} = \frac{173}{173+2} = \frac{173}{175} \approx 0.9573 = 95.73\% \quad (3.7)$$

***F*1-міра**:

$$F1_{train} = 2 * \frac{P*R}{P+R} = 2 * \frac{0.982*0.97}{0.982+0.97} = 2 * \frac{0.95254}{1.952} \approx 0.9745 = 97.45\% \quad (3.8)$$

$$F1_{train} = 2 * \frac{0.9886*0.9301}{0.9886+0.9301} = 2 * \frac{0.9199}{1.9187} \approx 0.9745 = 95.73\% \quad (3.8)$$

- Точність: 97.45%; 95.73%
- Модель добре справляється зі спам-повідомленнями (рис. 3.6), зберігаючи при цьому високу продуктивність.

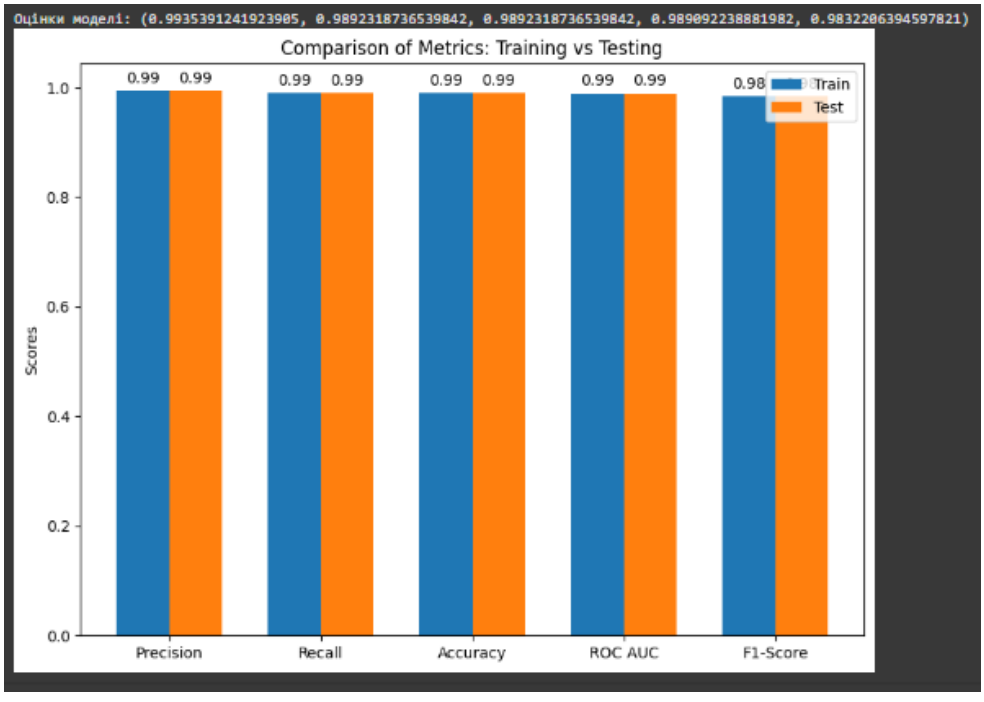

Рисунок 3.6 – Графік ефективності моделі машинного навчання

ROC AUC:

Високий *ROC AUC* підтверджує здатність моделі відокремлювати спам і не-спам (рис. 3.7).

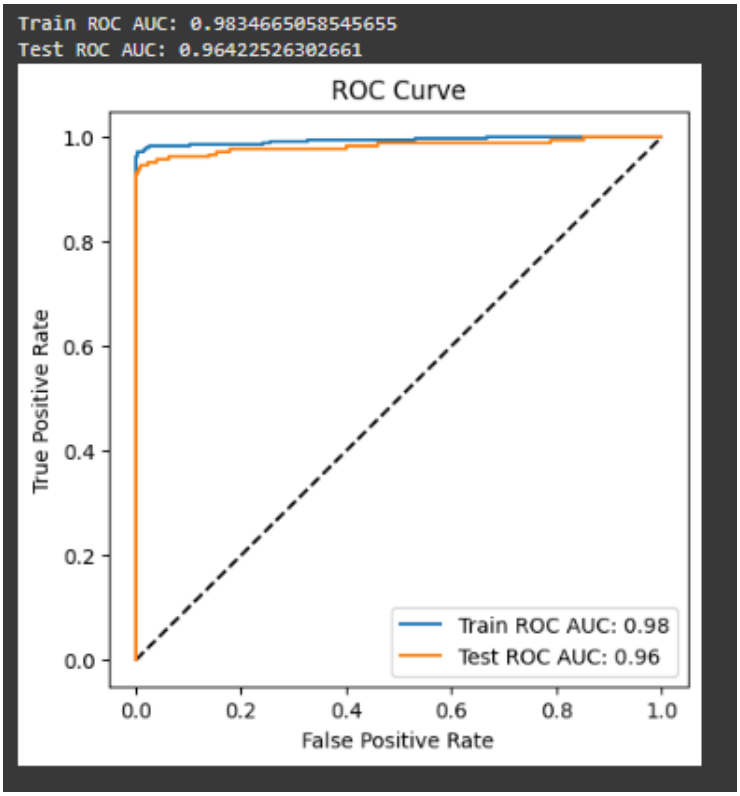

Рисунок 3.7 – Графік тестового та тренувального набору на ROC curve

- Для тренувальної вибірки: 98.92%
- Для тестової вибірки: 98.32%

4. Аналіз помилок

Результати матриці сплутаності демонструють (рис. 3.8), що більшість помилок моделі пов'язані з помилковою класифікацією не-спамових повідомлень як спаму (помилки типу *I*).

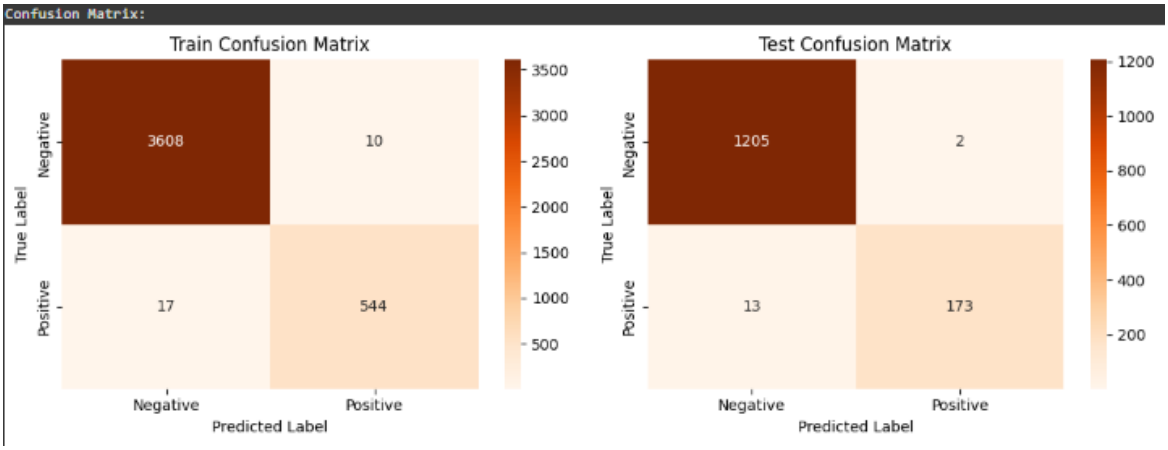

Рисунок 3.8 – Матриця сплутаності

Матриця сплутаності (*Test set*):

***Train*:**

- Істинно позитивні (*TP*): 544.
- Істинно негативні (*TN*): 3608.
- Хибно позитивні (*FP*): 10.
- Хибно негативні (*FN*): 17.

***Test*:**

- Істинно позитивні (*TP*): 173.
- Істинно негативні (*TN*): 1205.
- Хибно позитивні (*FP*): 2.
- Хибно негативні (*FN*): 13.

Модель краще справляється з розпізнаванням звичайних повідомлень, але складні спамові повідомлення все ще можуть бути помилково класифіковані.

5. Практичне застосування

Розроблена аналітична система успішно визначає спам, завдяки набору даних, використовуючи сучасні методи обробки тексту та машинного навчання. Її результати можуть бути інтегровані в реальні системи кіберзахисту для фільтрації шкідливого контенту, наприклад:

- Автоматизація фільтрації вхідної електронної пошти - забезпечення безпеки користувачів від шкідливих посилань та фішингових атак.

- Оцінка ризику під час спам-кампаній - Визначення та блокування джерел масової розсилки спаму.
- Використання у кіберполіції -Допомога у блокуванні ресурсів, що розсилають шкідливі повідомлення.

6. Приклад використання системи.

Було виконано перевірку роботи системи на реальному повідомленні (рис. 3.9):

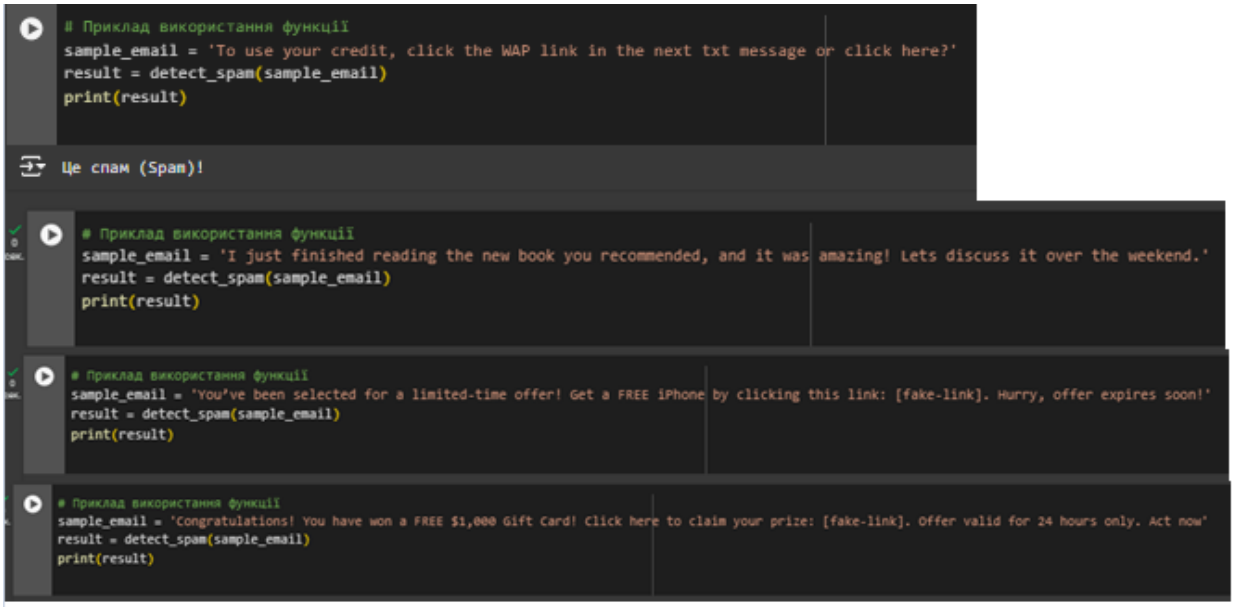

Рисунок 3.9 – Результатит текстового повідомлення та вивід результату.

sample_email = 'To use your credit, click the WAP link in the next txt message or click here?'

result = detect_spam(sample_email)

print(result)

Система правильно ідентифікує підозрілі повідомлення як спам, що підтверджує її ефективність.

3.5. Висновки до третього розділу

У третьому розділі були вирішені завдання, зазначені у вступі, що стосуються розробки та впровадження аналітичної системи для оцінки ефективності фільтрації спаму при кібератаках. Зокрема, була створена система, яка не лише фільтрує спам,

а й оцінює ефективність фільтрації за допомогою різноманітних метрик, таких як точність, чутливість, специфічність, *F*1-міра та *ROC AUC*. Система використовує сучасні методи машинного навчання, що дозволяють адаптувати її до нових типів спаму та оперативно реагувати на зміни в умовах кіберзагроз.

Завдання, виконані в розділі:

- Розробка аналітичної системи для оцінки ефективності фільтрації спаму - створена система, яка оцінює ефективність фільтрації за допомогою стандартних метрик та надає візуалізацію результатів.
- Оцінка ефективності фільтрації спаму при кібератаках — проведено оцінювання різних моделей фільтрації на основі метрик точності, чутливості, специфічності, а також *ROC AUC*, що дозволяє отримати всебічне уявлення про продуктивність системи.
- Інтеграція результатів у реальні системи кіберзахисту — розроблена система успішно визначає спам і може бути використана для автоматизації фільтрації спаму в реальних умовах, наприклад, в корпоративних системах електронної пошти або в кіберполіції.

Результати цього розділу показують, що розроблена аналітична система здатна ефективно оцінювати та покращувати фільтрацію спаму в умовах кібератак. Завдяки інтеграції машинного навчання та візуалізації результатів система може не лише фільтрувати спам, а й надавати необхідні дані для подальшої оптимізації фільтраційних алгоритмів. Це забезпечує високий рівень захисту від спаму та кіберзагроз у реальному часі.

ВИСНОВКИ

Проведені дослідження показали, що проблема ефективної фільтрації спаму залишається актуальною через постійне збільшення кіберзагроз, обсягів небажаних повідомлень та їхнього впливу на інформаційну безпеку. Розробка сучасних аналітичних систем для оцінки ефективності методів фільтрації є необхідною умовою забезпечення стабільної роботи інформаційних систем.

При виконанні кваліфікаційної роботи були розглянуті актуальні аспекти виявлення та фільтрації спаму, що дозволило сформувати системний підхід до оцінки ефективності різних методів.

Для досягнення мети дослідження було виконано наступні завдання:

1. Проведено аналіз сучасних методів виявлення та фільтрації спаму, таких як підходи на основі ключових слів, алгоритми ймовірнісного аналізу, машинного навчання та поведінкові моделі. Визначено їх сильні та слабкі сторони.
2. Визначено ключові показники для оцінки системи фільтрації спаму, включаючи точність, повноту, специфічність, F1-міру та швидкість обробки даних.
3. Реалізовано функціонал аналітичної системи, що дозволяє виконувати порівняльний аналіз ефективності фільтрації спаму.
4. Проведено тестування розробленої системи на реальних наборах даних із платформи Kaggle. Отримані результати підтвердили її здатність до точного виявлення спаму в умовах реальних загроз.
5. Оцінено вплив розробленої системи на технічні показники безпеки інформаційних систем.

У результаті виконання кваліфікаційної роботи досягнуто мети, яка полягала у створенні аналітичної системи для оцінки ефективності методів фільтрації спаму. Розроблена система сприяє покращенню захисту інформаційних систем завдяки можливості комплексного аналізу методів та адаптації до нових типів загроз.

Також під час виконання кваліфікаційної роботи було опубліковано роботу «Порівняльний аналіз популярних механізмів повнотекстового пошуку» [36]. В цьому тексті проаналізовані популярні механізми повнотекстового пошуку, такі як

Elasticsearch, Apache Solr і Apache Lucene. У ньому розглядаються їхні переваги (швидкість, масштабованість, підтримка складних запитів) та недоліки (високі вимоги до ресурсів, складність налаштування). Текст також пропонує рекомендації щодо вибору кожного механізму для конкретних завдань, таких як великі бізнес-проєкти, веб-додатки чи створення спеціалізованих рішень для пошуку.

Подальші дослідження можуть бути спрямовані на вдосконалення алгоритмів системи для підвищення її адаптивності та інтеграції у реальні середовища кіберзахисту. Це дозволить значно зменшити ризики, пов’язані зі спамом, і підвищити загальний рівень інформаційної безпеки.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. 2ip.ua. Що таке спам? URL: https://2ip.ua/ua/blog/spam (дата звернення: 16.11.2024).

2. Sebastiani F. Machine learning in automated text categorization // ACM Computing Surveys (CSUR). 2002. Vol. 34, No. 1. P. 1–47.

3. SendPulse. Що таке SMS-реклама? URL: https://sendpulse.ua/support/glossary/sms-advertising (дата звернення: 16.11.2024).

4. SendPulse. SMS-реклама: визначення та переваги. URL: https://sendpulse.ua/support/glossary/sms-advertising (дата звернення: 16.11.2024).

5. MaxNet. Фішингові сайти: як розпізнати шахрайський сайт? URL: https://maxnet.ua/blog/fishingovye-sayty-kak-raspoznat-moshennicheskiy-sayt/ (дата звернення: 16.11.2024).

6. VUE. Антиреклама. URL: https://vue.gov.ua/Антиреклама (дата звернення: 16.11.2024).

7. Wikipedia. Нігерійські листи. URL: https://uk.wikipedia.org/wiki/Нігерійські_листи (дата звернення: 16.11.2024).

8. Wikipedia. DoS-атака. URL: https://uk.wikipedia.org/wiki/DoS-атака (дата звернення: 18.11.2024).

9. Wikipedia. Дипфейк-фішинг. URL: https://uk.wikipedia.org/wiki/Дипфейк-фішинг (дата звернення: 18.11.2024).

10. Best Free Soft. Спам: види спаму і боротьба зі спамом. URL: https://best-free-soft.at.ua/publ/spam_vidi_spamu_i_borotba_zi_spamom/1-1-0-33 (дата звернення: 18.11.2024).

11. ESET. Мережевий черв. URL: https://www.eset.com/ua/support/information/entsiklopediya-ugroz/setevoy-cherv/?srsltid=AfmBOophG5onmEcF--L1ZIUMeXv5Q85RFXnAbXM8TT7lphRZ669GwW7r (дата звернення: 18.11.2024).

12. UA5.org. Основні ознаки спаму і фішингу. URL: https://ua5.org/protect/2217-osnovni-oznaky-spamu-j-fishyngu.html (дата звернення: 18.11.2024).

13. ESET. Антиспам. URL: https://www.eset.com/ua/support/information/entsiklopediya-ugroz/antyspam (дата звернення: 18.11.2024).

14. Bouncer. Методи обходу спам-фільтрів. URL: https://www.usebouncer.com/uk/методи-обходу-спам-фільтра/ (дата звернення: 19.11.2024).

15. GSminfo. Експерти розповіли, для чого і як саме працює фільтр спаму в електронній пошті. URL: https://gsminfo.com.ua/77518-eksperty-rozpovily-dlya-chogo-i-yak-same-praczyuye-filtr-spamu-v-elektronnij-poshti.html (дата звернення: 19.11.2024).

16. Wikipedia. Байєсова фільтрація спаму. URL: https://uk.wikipedia.org/wiki/Баєсова_фільтрація_спаму#:~:text=Naive%20Bayes%20spam%20filtering)%20—%20метод,на%20пряме%20використання%20теореми%20Баєса (дата звернення: 20.11.2024).

17. Wikipedia. Наївний баєсів класифікатор. URL: https://uk.wikipedia.org/wiki/Наївний_баєсів_класифікатор (дата звернення: 20.11.2024).

18. Wikipedia. Машинне навчання. URL: https://uk.wikipedia.org/wiki/Машинне_навчання (дата звернення: 20.11.2024).

19. Tuthost. Поняття спам і методи протидії. URL: https://tuthost.ua/uk/blog/ponyattya-spam-a-ta-metodi-protidiyi- (дата звернення: 16.11.2024).

20. Wikipedia. Медіа-маніпуляція. URL: https://uk.wikipedia.org/wiki/Медіа-маніпуляція (дата звернення: 20.11.2024).

21. Wikipedia. Інформаційний шум. URL: https://uk.wikipedia.org/wiki/Інформаційний_шум (дата звернення: 20.11.2024).

22. Wikipedia. Спам. URL: https://uk.wikipedia.org/Спам (дата звернення: 20.11.2024).

23. Termin. Спам: механізм роботи спаму, як це працює. URL: https://termin.in.ua/spam/#Mehanizm_roboti_spamu_ak_ce_pracue (дата звернення: 20.11.2024).

24. Wikipedia. Аналіз поведінки користувачів. URL: https://uk.wikipedia.org/wiki/Аналіз_поведінки_користувачів (дата звернення: 20.11.2024).

25. Nextcloud. Двофакторна автентифікація. URL: https://docs.nextcloud.com/server/latest/user_manual/uk/user_2fa.html (дата звернення: 20.11.2024).

26. Wikipedia. DNS spoofing. URL: https://uk.wikipedia.org/wiki/DNS_spoofing (дата звернення: 20.11.2024).

27. Wikipedia. Байєсова фільтрація спаму. URL: https://uk.wikipedia.org/wiki/Баєсова_фільтрація_спаму (дата звернення: 20.11.2024).

28. GeeksforGeeks. Text preprocessing for NLP tasks. URL: https://www.geeksforgeeks.org/text-preprocessing-for-nlp-tasks (дата звернення: 20.11.2024).

29. Finxter. 4 best ways to strip HTML tags from a Python string. URL: https://blog.finxter.com/4-best-ways-to-strip-html-tags-from-a-python-string (дата звернення: 20.11.2024).

30. Spamhaus. The Policy Blocklist. URL: https://www.spamhaus.org/resource-hub/dnsbl/the-policy-blocklist-what-is-it-and-why-should-you-be-on-it/#lessfont-style%22vertical-align:-inherit%22greaterlessfont-style%22vertical-align:-inherit%22greaterkoli-bulo-vvedeno-chornij-spisok-politikilessfontgreaterlessfontgreater (дата звернення: 21.11.2024).

31. Eset. Whitelist. URL: https://help.eset.com/glossary/uk-UA/whitelist.html (дата звернення: 21.11.2024).

32. Ukraine.com.ua. Web servers. Apache. URL: https://www.ukraine.com.ua/uk/wiki/hosting/web-servers/apache/htaccess/access-restrict/ (дата звернення: 21.11.2024).

33. SendPulse. Email authentication. URL: https://sendpulse.ua/knowledge-base/email-service/additional/email-authentication (дата звернення: 21.11.2024).

34. Wikipedia. Гіперпосилання. URL: https://uk.wikipedia.org/wiki/Гіперпосилання (дата звернення: 21.11.2024).

35. Wikipedia. Точність. URL: https://uk.wikipedia.org/wiki/Точність (дата звернення: 21.11.2024).

36. Ковальчук М.В., Струзік В.А. Порівняльний аналіз популярних механізмів повнотекстового пошуку // Перша міжнародна науково-практична конференція «Штучний інтелект та інформаційні технології», 3–4 червня 2024 р., Київ : Національний університет харчових технологій, 2024. С. 123–124.

37. UChoose. Комп'ютерна пропаганда. URL: https://uchoose.uacrisis.org/kompyuterna-propaganda (дата звернення: 22.11.2024).

38. Chavez A. F., Liu B. A survey of spam filtering techniques // IEEE Communications Magazine. 2003. Vol. 41, No. 8. P. 33–38.

39. Sahami M., Dumais S. T., Heckerman D. A Bayesian approach to filtering junk email // Proceedings of the AAAI-98 Workshop on Learning for Text Categorization. 1998. P. 55–62.

40. Yang Y., Pedersen J. O. A comparative study on feature selection in text categorization // Proceedings of the Fourteenth International Conference on Machine Learning (ICML-97). 1997. P. 412–420.

41. Koller D., Witten I. H. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2007. 758 p.

42. Joulin A., Grave E., Mikolov T., Rousset M. Bag of Tricks for Efficient Text Classification. 2017.

43. Patel H., Patel P. Spam detection and filtering techniques: A survey // International Journal of Computer Science and Information Technologies. 2016. Vol. 7, No. 5. P. 2081–2086.

44. Bekkerman R., Allan J. Using Out-of-Context Words for Spam Filtering // Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2004. P. 377–384.

45. Koller D., Witten I. H. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2007. 758 p.

46. Kaggle. Spam dataset. URL: https://www.kaggle.com/datasets/tmehul/spamcsv (дата звернення: 14.11.2024).

47. Брюс Дж. Вступ до машинного навчання. — К.: Книга, 2018. — 320 с.

48. Кохан І. О. Кібербезпека: основи та практичні аспекти. — Х.: ХНУ, 2019. — 410 с.

49. Гарнер С. Безпека в Інтернеті: від теорії до практики. — К.: Видавництво "Техніка", 2017. — 550 с.

50. Кунц Д. Системи обробки даних. — К.: Видавничий центр "Наука", 2020. — 512 с.

51. Sakari K. Data Privacy and Spam Control in Machine Learning Systems. — London: Springer, 2018. — 350 p.

52. Raghu M. Foundations of Cybersecurity.-New York: Wiley, 2019. — 330 p.

53. Bernstein P., Robbins J. Network Security Essentials: Applications and Standards. — Pearson, 2017. — 400 p.

54. Zhou H., Liu C. Deep Learning in Cybersecurity. — San Francisco: Elsevier, 2020. — 242 p.

55. Hodge P. Cyber Attack Defense for the Internet of Things. — Springer, 2021. — 289 p.

56. Dummett J. Principles of Network Security. — London: Routledge, 2020. — 510 p.

57. Cheswick W. R., Bellovin S. M. Firewalls and Internet Security: Repelling the Wily Hacker. — Addison-Wesley, 2003. — 496 p.

58. Evans D. M. Network Security Monitoring. — Sebastopol: O'Reilly Media, 2019. — 408 p.

59. Solomon M. Building a Secure & Privacy-Focused IoT Network. — Cambridge: MIT Press, 2021. — 370 p.

60. Meier J., Mackman A., Dunner B. Improving Web Application Security: Threats and Countermeasures. — Microsoft Press, 2003. — 500 p.

61. Stallings W. Cryptography and Network Security: Principles and Practice. — Pearson, 2020. — 840 p.

62. Fonseca J. Cybersecurity in Internet of Things: Challenges and Solutions. — Springer, 2019. — 320 p.

63. Tavakoli S., Razzaghzadeh N. Spam Filtering Using Deep Learning Models. // International Journal of Information Security Science. 2018. Vol. 7, No. 3. P. 240–252.

64. Marti S., Garcia-Molina H. Taxonomy of trust: Categorizing P2P reputation systems. // Computer Networks. 2006. Vol. 50, No. 4. P. 472–484.

65. Garfinkel S., Spafford G. Practical UNIX & Internet Security. — Sebastopol: O'Reilly Media, 2003. — 960 p.

66. Dowd M., McDonald J., Schuh J. The Art of Software Security Assessment: Identifying and Preventing Software Vulnerabilities. - Addison-Wesley, 2006. -1200 p.

67. Bishop M. Computer Security: Art and Science. — Addison-Wesley, 2002. — 1080 p.

68. Parker D. Fighting Computer Crime. — Wiley, 1998. — 688 p.

69. Schneier B. Applied Cryptography: Protocols, Algorithms, and Source Code in C. — Wiley, 1996. — 784 p.

70. Peterson J., Dell M. Big Data Analytics in Cybersecurity. — CRC Press, 2018. — 270 p.

71. Goldstein J., Tapp J. Intelligent Systems for Cybersecurity. — Springer, 2017. — 360 p.

72. Yan Z. Cybersecurity Analytics and Operations. — Wiley, 2020. — 400 p.

73. Shon T., Moon J. Big Data and Cybersecurity Challenges. // Journal of Information Security. 2019. Vol. 9, No. 4. P. 200–216.

74. Mitnick K., Simon W. The Art of Intrusion: The Real Stories Behind the Exploits of Hackers, Intruders & Deceivers. — Wiley, 2005. — 288 p.

75. Allen J. The CERT Guide to Insider Threats. — Addison-Wesley, 2012. — 360 p.

76. Alazab M., Tang J. Cybercrime and Cybersecurity in the Global South. — Springer, 2019. — 250 p.

77. Brennan S., Warren G. Cyber Threat Modeling and Mitigation. — CRC Press, 2021. — 320 p.

78. Maillart S., Pasquini E. Algorithmic Approaches in Spam Detection and Filtering. // ACM Transactions on Information Systems. 2021. Vol. 39, No. 1. P. 1–28.

79. Singh A., Arora N. Machine Learning for Cybersecurity Applications. — Springer, 2022. — 300 p.

80. Baker A., Willis P. Advances in Network Security: Best Practices. — CRC Press, 2019. — 280 p.

81. Monroe R. Real-Time Cybersecurity Analytics. — Elsevier, 2021. — 310 p.

82. Cohen J., Smith R. Practical Techniques in Information Security. — McGraw-Hill, 2020. — 400 p.

ДОДАТКИ

Додаток А. Код програми

```
# Імпорт бібліотек
# Імпорт Numpy та Pandas
import numpy as np
import pandas as pd

# Імпорт інструментів для візуалізації
import matplotlib.pyplot as plt
import seaborn as sns

# Імпорт бібліотек для оцінки метрик
from sklearn.metrics import confusion_matrix, accuracy_score,
precision_score, recall_score, f1_score, roc_auc_score, roc_curve,
classification_report

# Бібліотека для створення хмар слів
from wordcloud import WordCloud, STOPWORDS

# Бібліотека для попередньої обробки текстових даних
from sklearn.feature_extraction.text import CountVectorizer

# Імпорт бібліотек для поділу даних на тренувальні і тестові набори
from sklearn.model_selection import train_test_split

# Бібліотека для реалізації моделей машинного навчання
from sklearn.naive_bayes import MultinomialNB

# Імпорт класу Pipeline зі scikit-learn
from sklearn.pipeline import Pipeline

# Бібліотека для ігнорування попереджень
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline

# Завантаження датасету з локального файлу
df = pd.read_csv("spam.csv", encoding='ISO-8859-1')

# Перший погляд на датасет
# Перегляд перших 5 рядків датасету
df.head()

# Кількість рядків та стовпців у датасеті
# Перевірка кількості рядків і стовпців датасету за допомогою shape
print("Кількість рядків: ", df.shape[0])
print("Кількість стовпців: ", df.shape[1])

# Інформація про датасет
```

```
# Перевірка структури датасету за допомогою info()
df.info()

# Кількість дублікатів у датасеті
dup = df.duplicated().sum()
print(f'Кількість дубльованих рядків: {dup}')
# Перевірка пропущених значень
df.isnull().sum()
# Назви стовпців датасету
df.columns
# Статистичний опис датасету (включаючи всі стовпці)
df.describe(include='all').round(2)

# Перевірка кількості унікальних значень для кожної змінної за допомогою
циклу
for i in df.columns.tolist():
     print("Кількість унікальних значень у стовпці", i, "становить",
df[i].nunique())

# Зміна назв стовпців v1 та v2 на Category та Message
df.rename(columns={"v1": "Category", "v2": "Message"}, inplace=True)

# Видалення всіх стовпців без імен (які містять багато пропущених значень)
df.drop(columns={'Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'}, inplace=True)

# Створення бінарного стовпця 'Spam': 1 для 'spam' і 0 для 'ham', на основі
стовпця 'Category'.
df['Spam'] = df['Category'].apply(lambda x: 1 if x == 'spam' else 0)

# Оновлений датасет
df.head()

# Графік - 1: Кругова діаграма для розподілу спам-повідомлень та не спам-
повідомлень
spread = df['Category'].value_counts()
plt.rcParams['figure.figsize'] = (5, 5)

# Встановлення міток
spread.plot(kind='pie', autopct='%1.2f%%', cmap='Set1')
plt.title(f'Розподіл спам-повідомлень і не спам-повідомлень')

# Виведення графіка
plt.show()

# Розділення спам-повідомлень
df_spam = df[df['Category'] == 'spam'].copy()

# Графік - 2: Візуалізація WordCloud для найбільш вживаних слів у спам-
повідомленнях
# Створення рядка для збереження всіх слів
comment_words = ''
```

```
# Видалення стоп-слів
stopwords = set(STOPWORDS)

# Прохід через стовпець повідомлень
for val in df_spam.Message:

    # Перетворення кожного значення в рядок
    val = str(val)

    # Розбиття значення на токени
    tokens = val.split()

    # Перетворення кожного токену на малу літеру
    for i in range(len(tokens)):
        tokens[i] = tokens[i].lower()

    comment_words += " ".join(tokens) + " "

# Встановлення параметрів
wordcloud = WordCloud(width=1000, height=500,
                background_color='white',
                stopwords=stopwords,
                min_font_size=10,
                max_words=1000,
                colormap='gist_heat_r').generate(comment_words)

# Встановлення заголовка
plt.figure(figsize=(6, 6), facecolor=None)
plt.title('Найбільш вживані слова у спам-повідомленнях', fontsize=15,
pad=20)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad=0)

# Виведення графіка
plt.show()

# Розділення даних на навчальну та тестову вибірки
X_train, X_test, y_train, y_test = train_test_split(df.Message, df.Spam,
test_size=0.25)

def evaluate_model(model, X_train, X_test, y_train, y_test):
    '''Функція приймає модель, X train, X test, y train, y test
    і виконує наступні дії:
    1. Навчає модель на тренувальних даних.
    2. Створює прогнози для тренувальних та тестових даних.
    3. Виводить ROC-AUC score для тренувальних та тестових даних.
    4. Виводить та будує ROC-криву.
    5. Виводить матрицю сплутаності для тренувальних та тестових даних.
    6. Виводить звіт про класифікацію для тренувальних та тестових даних.
    7. Виводить важливість ознак, якщо модель її підтримує.
    8. Повертає список з наступними оцінками:
```

```
      recall_train, recall_test, acc_train, acc_test, roc_auc_train,
roc_auc_test, F1_train, F1_test
    '''

    # Навчання моделі на тренувальних даних
    model.fit(X_train, y_train)

    # Прогнози для тренувальних та тестових даних
    y_pred_train = model.predict(X_train)
    y_pred_test = model.predict(X_test)
    pred_prob_train = model.predict_proba(X_train)[:,1]
    pred_prob_test = model.predict_proba(X_test)[:,1]

    # Обчислення ROC AUC score
    roc_auc_train = roc_auc_score(y_train, y_pred_train)
    roc_auc_test = roc_auc_score(y_test, y_pred_test)
    print("\nTrain ROC AUC:", roc_auc_train)
    print("Test ROC AUC:", roc_auc_test)

    # Побудова ROC кривої
       fpr_train, tpr_train, thresholds_train = roc_curve(y_train,
pred_prob_train)
        fpr_test, tpr_test, thresholds_test = roc_curve(y_test,
pred_prob_test)
    plt.plot([0,1],[0,1],'k--')
         plt.plot(fpr_train, tpr_train, label="Train ROC AUC:
{:.2f}".format(roc_auc_train))
          plt.plot(fpr_test, tpr_test, label="Test ROC AUC:
{:.2f}".format(roc_auc_test))
    plt.legend()
    plt.title("ROC Curve")
    plt.xlabel("False Positive Rate")
    plt.ylabel("True Positive Rate")
    plt.show()

    # Обчислення матриці сплутаності
    cm_train = confusion_matrix(y_train, y_pred_train)
    cm_test = confusion_matrix(y_test, y_pred_test)

    fig, ax = plt.subplots(1, 2, figsize=(11,4))

    print("\nConfusion Matrix:")
         sns.heatmap(cm_train, annot=True, xticklabels=['Negative',
'Positive'], yticklabels=['Negative', 'Positive'], cmap="Oranges",
fmt='.4g', ax=ax[0])
    ax[0].set_xlabel("Predicted Label")
    ax[0].set_ylabel("True Label")
    ax[0].set_title("Train Confusion Matrix")

   sns.heatmap(cm_test, annot=True, xticklabels=['Negative', 'Positive'],
yticklabels=['Negative', 'Positive'], cmap="Oranges", fmt='.4g', ax=ax[1])
    ax[1].set_xlabel("Predicted Label")
```

```
    ax[1].set_ylabel("True Label")
    ax[1].set_title("Test Confusion Matrix")

    plt.tight_layout()
    plt.show()

    # Обчислення звіту про класифікацію
         cr_train    =    classification_report(y_train,    y_pred_train,
output_dict=True)
    cr_test = classification_report(y_test, y_pred_test, output_dict=True)
    print("\nTrain Classification Report:")
    crt = pd.DataFrame(cr_train).T
    print(crt.to_markdown())
      #  sns.heatmap(pd.DataFrame(cr_train).T.iloc[:,  :-1],  annot=True,
cmap="Blues")
    print("\nTest Classification Report:")
    crt2 = pd.DataFrame(cr_test).T
    print(crt2.to_markdown())
      #  sns.heatmap(pd.DataFrame(cr_test).T.iloc[:,  :-1],  annot=True,
cmap="Blues")

    precision_train = cr_train['weighted avg']['precision']
    precision_test = cr_test['weighted avg']['precision']

    recall_train = cr_train['weighted avg']['recall']
    recall_test = cr_test['weighted avg']['recall']

    acc_train = accuracy_score(y_true = y_train, y_pred = y_pred_train)
    acc_test = accuracy_score(y_true = y_test, y_pred = y_pred_test)

    F1_train = cr_train['weighted avg']['f1-score']
    F1_test = cr_test['weighted avg']['f1-score']

      model_score  =  [precision_train,  precision_test,  recall_train,
recall_test, acc_train, acc_test, roc_auc_train, roc_auc_test, F1_train,
F1_test ]
    return model_score

# Моделювання ML - 1 Реалізація
# Створення машинного навчання через pipeline з scikit-learn, поєднуючи
векторизацію тексту (CountVectorizer)
#  та класифікацію за допомогою наївного баєсового класифікатора для
виявлення спаму в електронних листах.
clf = Pipeline([
    ('vectorizer', CountVectorizer()),  # Крок 1: Перетворення текстових
даних
    ('nb', MultinomialNB())  # Крок 2: Класифікація за допомогою Наївного
Баєса
])

# Модель навчається (fit) та робить прогнози в функції evaluate_model
```

```
# Visualizing evaluation Metric Score chart
MultinomialNB_score = evaluate_model(clf, X_train, X_test, y_train,
y_test)

# Опис функції для системи виявлення спаму в електронних листах
def detect_spam(email_text):

    # Робимо прогноз за допомогою класифікатора
    prediction = clf.predict([email_text])

    if prediction == 0:
        return "Це не спам (Ham)!"
    else:
        return "Це спам (Spam)!"

# Приклад використання функції
sample_email = 'To use your credit, click the WAP link in the next txt
message or click here?'
result = detect_spam(sample_email)
print(result)
```

Додаток Б. Скріншоти роботи програми

```
[5]  # Перевірка кількості унікальних значень для кожної змінної за допомогою циклу
     for i in df.columns.tolist():
         print("Кількість унікальних значень у стовпці", i, "становить", df[i].nunique())
```

```
Кількість унікальних значень у стовпці v1 становить 2
Кількість унікальних значень у стовпці v2 становить 5169
Кількість унікальних значень у стовпці Unnamed: 2 становить 43
Кількість унікальних значень у стовпці Unnamed: 3 становить 10
Кількість унікальних значень у стовпці Unnamed: 4 становить 5
```

```
[6]  # Зміна назв стовпців v1 та v2 на Category та Message
     df.rename(columns={"v1": "Category", "v2": "Message"}, inplace=True)
```

```
[7]  # Видалення всіх стовпців без імен (які містять багато пропущених значень)
     df.drop(columns={'Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'}, inplace=True)
```

```
[8]  # Створення бінарного стовпця 'Spam': 1 для 'spam' і 0 для 'ham', на основі стовпця 'Category'.
     df['Spam'] = df['Category'].apply(lambda x: 1 if x == 'spam' else 0)
```

```
[9]  # Оновлений датасет
     df.head()
```

	Category	Message	Spam
0	ham	Go until jurong point, crazy.. Available only ...	0
1	ham	Ok lar... Joking wif u oni...	0
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	1
3	ham	U dun say so early hor... U c already then say...	0
4	ham	Nah I don't think he goes to usf, he lives aro...	0

Рисунок Б.1 – Створення бінарного стовпця та перевірка стовпців на помилки

Рисунок Б.2 – Найбільш вживані слова у спам-повідомленнях

```
[15] # Моделювання ML - 1 Реалізація
     # Створення машинного навчання через pipeline з scikit-learn, поєднуючи векторизацію тексту (CountVectorizer)
     # та класифікацію за допомогою наївного баєсового класифікатора для виявлення спаму в електронних листах.
     clf = Pipeline([
         ('vectorizer', CountVectorizer()),  # Крок 1: Перетворення текстових даних
         ('nb', MultinomialNB())  # Крок 2: Класифікація за допомогою Наївного Баєса
     ])

     # Модель навчається (fit) та робить прогнози в функції evaluate_model
```

```
# Visualizing evaluation Metric Score chart
MultinomialNB_score = evaluate_model(clf, X_train, X_test, y_train, y_test)
```

Рисунок Б.3 – Виконане машинне навчання по набору даних

```
Train Classification Report:
|              |   precision |   recall |   f1-score |     support |
|:-------------|------------:|---------:|-----------:|------------:|
| 0            |     0.99531 | 0.997236 |   0.996272 | 3618        |
| 1            |     0.981949 | 0.969697 |   0.975785 |  561        |
| accuracy     |     0.993539 | 0.993539 |   0.993539 |    0.993539 |
| macro avg    |     0.98863 | 0.983467 |   0.986029 | 4179        |
| weighted avg |     0.993517 | 0.993539 |   0.993522 | 4179        |

Test Classification Report:
|              |   precision |   recall |   f1-score |     support |
|:-------------|------------:|---------:|-----------:|------------:|
| 0            |     0.989327 | 0.998343 |   0.993814 | 1207        |
| 1            |     0.988571 | 0.930108 |   0.958449 |  186        |
| accuracy     |     0.989232 | 0.989232 |   0.989232 |    0.989232 |
| macro avg    |     0.988949 | 0.964225 |   0.976132 | 1393        |
| weighted avg |     0.989226 | 0.989232 |   0.989092 | 1393        |
```

Рисунок Б.4 – Звіт по тренувальному та тестовому набору

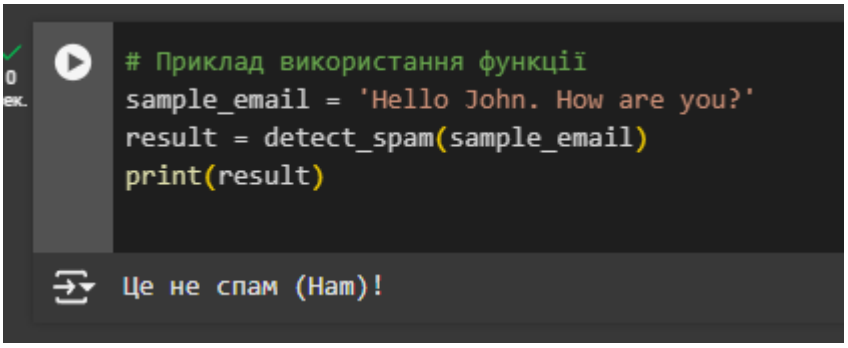

Рисунок Б.5 – Повторне використання прогнозування спаму в повідомленні

Додаток В. Схема основних етапів створення і розповсюдження спаму

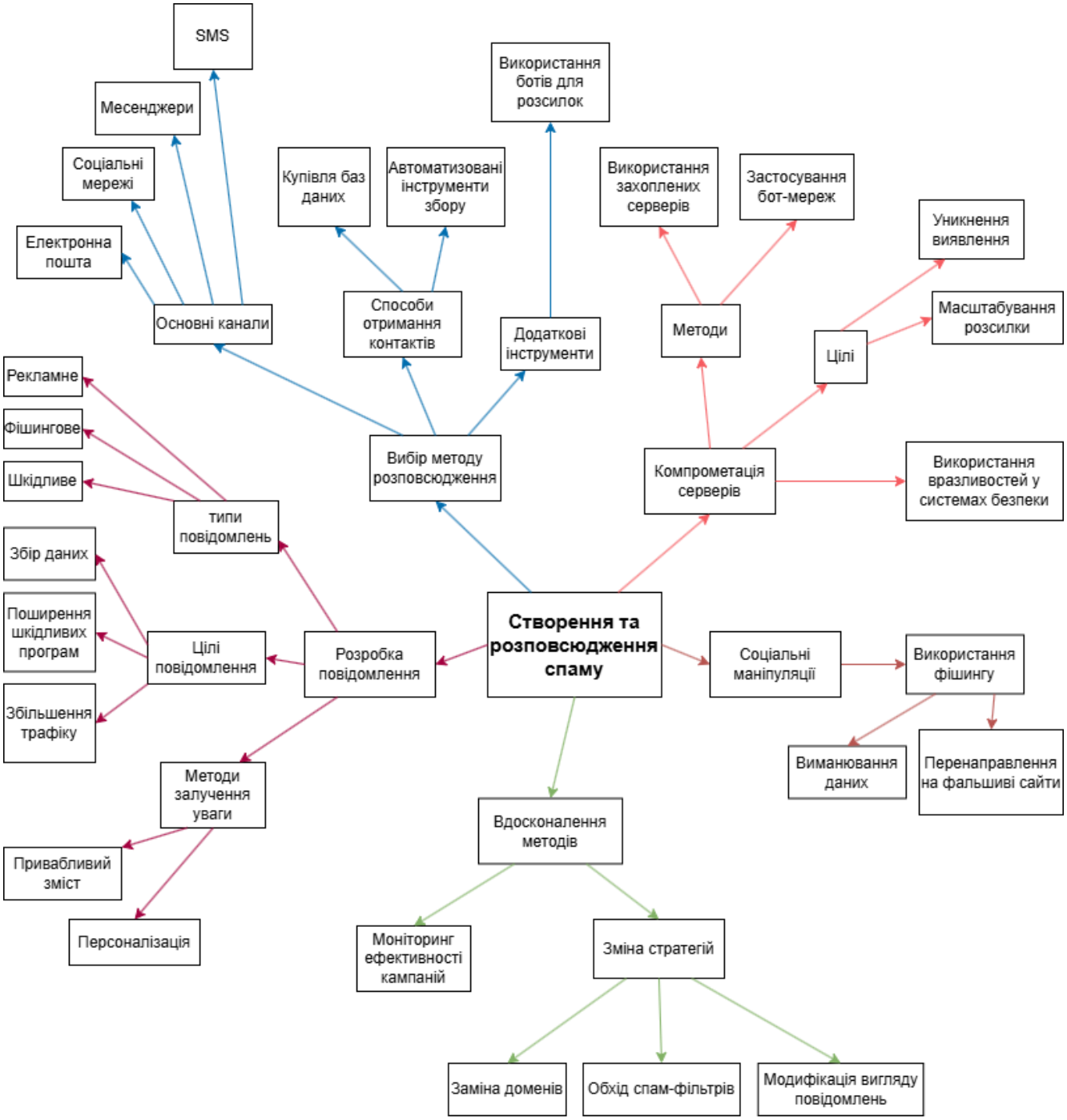

Рисунок В.1 – Схема основних етапів створення і розповсюдження спаму

Додаток Г. Представлення CSV файлу

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	v1,v2,,,																		
2	ham,"Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...",,,																		
3	ham,Ok lar... Joking wif u oni...,,,																		
4	spam,Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's,,,																		
5	ham,U dun say so early hor... U c already then say...,,,																		
6	ham,"Nah I don't think he goes to usf, he lives around here though",,,																		
7	spam,"FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, eJ1.50 to rcv",,,																		
8	ham,Even my brother is not like to speak with me. They treat me like aids patent.,,,																		
9	ham,As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune,,,																		
10	spam,WINNER!! As a valued network customer you have been selected to receivea eJ900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.,,,																		
11	spam,Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030,,,																		
12	ham,"I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.",,,																		
13	spam,"SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info",,,																		
14	spam,"URGENT! You have won a 1 week FREE membership in our eJ100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18",,,																		
15	ham,I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.,,,																		
16	ham,I HAVE A DATE ON SUNDAY WITH WILL!!,,,																		
17	spam,"XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL",,,																		
18	ham,Oh k...i'm watching here:),,,																		
19	ham,Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.,,,																		
20	ham,Fine if thateXs the way u feel. ThateXs the way its gota b,,,																		
21	spam,"England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/Mj1.20 POBOXox36504W45WQ 16+",,,																		
22	ham,Is that seriously how you spell his name?,,,																		
23	ham,I‰ыЧм going to try for 2 months ha ha only joking,,,																		
24	ham,So M_ pay first lar... Then when is da stock comin...,,,																		
25	ham,Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?,,,																		
26	ham,Fffffffff. Alright no way I can meet up with you sooner?,,,																		
27	ham,Just forced myself to eat a slice. I'm really not hungry tho. This sucks. Mark is getting worried. He knows I'm sick when I turn down pizza. Lol,,,																		
28	ham,Lol your always so convincing.,,,																		
29	ham,Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom's left over dinner ? Do you feel my Love ?,,,																		
30	ham,"I'm ｜ we're packing the car now, I'll let you know if there's room",,,																		
31	ham,Ahhh. Work. I vaguely remember that! What does it feel like? Lol,,,																		
32	ham,"Wait that's still not all that clear, were you not sure about me being sarcastic or that that's why x doesn't want to live with us",,,																		
33	ham,Yeah he got in at 2 and was v apologetic. n had fallen out and she was actin like spoilt child and he got caught up in that. Till 2! But we won't go there! Not doing too badly cheers. You? ,,,																		
34	ham,K tell me anything about you.,,,																		
35	ham,For fear of fainting with the of all that housework you just did? Quick have a cuppa,,,																		
36	spam,Thanks for your subscription to Ringtone UK your mobile will be charged eJ5/month Please confirm by replying YES or NO. If you reply NO you will not be charged,,,																		
37	ham,Yup... Ok i go home look at the timings then i msg M_ again... Xuhui going to learn on 2nd may too but her lesson is at 8am,,,																		
38	ham,"Oops, I'll let you know when my roommate's done",,,																		
39	ham,I see the letter B on my car,,,																		

Рисунок Г.1 – Представлення csv файлу датасету